



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

MARCELO DOS SANTOS

Uma releitura inferencial e análise de diagnósticos em modelos geoestatísticos

Recife

2022

MARCELO DOS SANTOS

Uma releitura inferencial e análise de diagnósticos em modelos geoestatísticos

Tese de doutorado apresentada ao Programa de Pós-graduação em Estatística do Departamento de Estatística e Probabilidade da Universidade Federal de Pernambuco como requisito parcial para obtenção do título de Doutor em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Profa. Dra. Fernanda De Bastiani

Coorientador (a): Prof. Dr. Miguel Angel Uribe Opazo

Recife

2022

Catálogo na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

S237r Santos, Marcelo dos
Uma releitura inferencial e análise de diagnósticos em modelos
geoestatísticos / Marcelo dos Santos. – 2022.
159 f.: il., fig., tab.

Orientadora: Fernanda de Bastiani.
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN,
Estatística, Recife, 2022.
Inclui referências e apêndices.

1. Estatística Aplicada. 2. Equações de estimação generalizada. 3. Família
Matérn. 4. Família Wendland. I. Bastiani, Fernanda de (orientadora). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2022 - 43

MARCELO DOS SANTOS

Uma releitura inferencial e análise de diagnósticos em modelos geoestatísticos

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Estatística.

Aprovada em: 24 de fevereiro de 2022.

BANCA EXAMINADORA

FERNANDA DE BASTIANI
UFPE

CRISTIANO FERRAZ
UFPE

GETÚLIO JOSÉ AMORIM DO AMARAL
UFPE

ORIETTA NICOLIS
UNIVERSIDAD ANDRES BELLO

MANUEL JESUS GALEA ROJAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Aos meus pais, por tudo que fizeram e me proporcionaram para que eu pudesse chegar até aqui. Doralice e José Augusto, a vocês todo o meu amor.

AGRADECIMENTOS

Primeiramente a Deus, que é o caminho, verdade e vida, por ter mantido esta chama acesa na direção desejada, mesmo quando ventos desfavoráveis tentaram apagá-la e me dando a coragem necessária para enfrentar os momentos difíceis.

A minha amada família: minha mãe Doralice dos Santos, meu pai José Augusto dos Santos, meus irmãos João José e Rogério Augusto por sempre acreditarem em mim e pela força ao longo de todos estes anos de minha formação. E minha esposa Gilayne Ferreira Santos, que em todos os momentos de dificuldade me transmitiu tranquilidade.

Aos meus amados filhos: João Henrique e Marcela Helena por serem minha fonte de inspiração.

A professora Dra. Fernanda De Bastiani, minha orientadora, pelos incentivos, ensinamentos, amizade e toda sua dedicação que, em todo tempo, demonstrou acreditar em minha capacidade e me guiando na direção correta para enfrentar os momentos difíceis.

Ao professor Dr. Miguel Angel Uribe Opazo, meu co-orientador, pelo apoio e disposição para auxiliar nas sugestões e correções.

Aos meus colegas de turma: César Diogo, José Iraponil, Maria Ioneris e Raimundo Cardoso pelas conversas, ensinamentos, críticas construtivas, apoio e carinho nos momentos difíceis. Obrigado por serem além de amigos, seres humanos incríveis.

A todos os colegas do programa de pós-graduação PPGEST pelo apoio, compreensão e carinho, pelas piadas e pela força durante minha caminhada até aqui.

Aos professores e funcionários do programa de pós-graduação PPGEST que contribuíram e fizeram parte para construção deste processo acadêmico.

Aos meus amigos e ex-alunos do IFBA-Campus Eunápolis pelas várias conversas e incentivo a sempre buscar e aprimorar meus conhecimentos.

Ao Instituto Federal de Educação, Ciência e Tecnologia da Bahia - IFBA, por conceder o auxílio de licença para capacitação.

Na tentativa de não omitir nenhum nome: a todos amigos de minha infância, especialmente do bairro de Salobrinho em Ilhéus (BA), dos colégios aos quais eu já fui professor: Professora Horizontina Conceição, Paulo Américo de Oliveira, Estadual do Basílio, SENAI e Escola Criativa.

RESUMO

O presente trabalho de tese, estuda situações em que os eventos de interesse apresentam dependência espacial entre as observações. Não obstante, estudos nesta área tenham tido notáveis avanços, técnicas relacionadas a eficiência da matriz de correlação espacial utilizada, bem como a escolha de tal, além de modelos de quasi-verossimilhança e distribuições para dados com excesso de zeros, ainda são pouco exploradas. Dito isto, desenvolvemos este trabalho em três etapas. Primeiro, utilizamos a família de funções de covariância Wendland e a comparamos com a família Matérn, para vários valores de seus parâmetros de suavização. Paralelamente, desenvolvemos as relações entre o alcance prático e o parâmetro de suporte e, estabelecemos expressões para mensurar o índice de dependência espacial para ambas as famílias. Além disso, como técnica de diagnóstico, são desenvolvidas medidas de influência local sob diferentes esquemas de perturbação. Como critério de identificação de observações potencialmente influentes, propomos utilizar níveis de referência obtidos a partir de reamostragem *Jackknife-after-Bootstrap*. Na segunda etapa, propomos um novo critério para seleção da matriz de correlação espacial de trabalho, baseado no condicionamento da matriz de variância-covariância *naïve*, a partir de modelos de quasi-verossimilhança e Equações de Estimação Generalizadas. Avaliamos a performance do método por um extenso estudo de simulação, utilizando as distribuições marginais da Normal, Poisson e Gama para dados espacialmente correlacionados. A especificação da estrutura de correlação é baseada em modelos de semivariogramas, utilizando as famílias Wendland, Matérn e Esférica. Os resultados revelaram que o critério proposto resultou em um melhor desempenho em comparação com os concorrentes disponíveis na literatura, com relação as taxas de acerto da verdadeira estrutura de correlação espacial dos dados simulados. Na terceira etapa, propomos um modelo geostatístico Gama aumentada com zero. Para acomodar a dependência espacial, incorporamos uma estrutura de quasi-verossimilhança e estimamos seus parâmetros utilizando o algoritmo *Expectation-Solution*. Além disso, sugerimos medidas para os resíduos e desenvolvemos medidas de influência local, sob os esquemas de perturbação ponderação de casos, perturbação na variável resposta e nas covariáveis. Por fim, acrescentamos que em todas as etapas um exemplo numérico foi analisado utilizando conjuntos de dados reais.

Palavras-chaves: equações de estimação generalizada; família Matérn; família Wendland; quasi-gama aumentada com zero; quasi-verossimilhança; variabilidade espacial.

ABSTRACT

This thesis work studies situations in which the events of interest present spatial dependence between the observations. Regardless of studies in this area have had notable advances, techniques related to the efficiency of the spatial correlation matrix used, as well as the choice of such, in addition to quasi-likelihood models and distributions for data with an excess of zeros, are still little explored. That said, we developed this work in three stages. In the first step, we use the Wendland family of covariance functions and compare it with the Matérn family, for various values of its smoothing parameters. In parallel, we developed the relationships between the practical range and the support parameter, and we established expressions to measure the spatial dependence index for both families. Furthermore, as a diagnostic technique, measurements of local influence are developed under different disturbance schemes. As a criterion for identifying potentially influential observations, we propose to use reference levels obtained from *Jackknife-after-Bootstrap* resampling. In the second step, we propose a new criterion for selecting the working spatial correlation matrix, based on conditioning the *naïve* variance-covariance matrix, based on quasi-likelihood models and Generalized Estimation Equations. We evaluated the performance of the method by an extensive study of simulations, using the marginal distributions of Normal, Poisson and Gamma for spatially correlated data. The specification of the correlation structure is based on semivariogram models, using the Wendland, Matérn and Spherica families. The results revealed that the proposed criterion resulted in better performance compared to the competitors available in the literature, regarding the success rates of the true spatial correlation structure of the simulated data. In the third step, we propose a zero-augmented Gamma geostatistical model. To accommodate the spatial dependence, we incorporate a quasi-likelihood structure and estimate its parameters using the *Expectation-Solution* algorithm. Furthermore, we suggest measures for residuals and develop measures of local influence, under the perturbation schemes weighting cases, perturbation in the response variable and in the covariates. Finally, we add that at all stages a numerical example was analyzed using real data sets.

Keywords: generalized estimation equations; Matérn family; quasi-gamma augmented with zero; quasi-likelihood; spatial variability; Wendland family.

LISTA DE FIGURAS

- Figura 1 – Modelos de correlação compatíveis para o caso $d = 2$. Para o modelo Matérn os parâmetros de suavização $\nu = 0,5$ e $\nu = 1,5$ com equivalência para Wendland $\kappa = 0$ e $\kappa = 1$. Equivalências entre os parâmetros de escala em (a) com $\alpha = 1/3$, fixo, $\beta = 0,667$ e $1,00$; em (b) $\alpha = 1/4,75$, fixo, $\beta = 1,038$ e $1,356$, calculados pela Equação (2.7). 32
- Figura 2 – Correlação de Wendland Generalizada para suavização $\kappa = 0$, em (a) e para $\kappa = 1$, em (b), para vários valores de ϕ_3 , com parâmetro que define a positividade da função de covariância $\mu = 5$ em ambos os casos. 36
- Figura 3 – Correlação de Wendland Generalizada para suavização $\kappa = 2$, em (a) e para $\kappa = 3$, em (b), para vários valores de ϕ_3 , com parâmetro que define a positividade da função de covariância $\mu = 5$ em ambos os casos. 37
- Figura 4 – Gráfico da relação entre a função de correlação e semivariância da Wendland Generalizada para suavização $\kappa = 3, 5$ (a). Funções de Semivariância para família de Wendland Generalizada e Matérn com suavização $\nu = 3, 5$. Com alcances prático aproximadamente iguais a $0,82$ e $0,76$ ($\phi_3 = 1,6$ e $0,11$) para Wendland e Matérn, respectivamente. Em que $g(\cdot)$ é função que varia para cada caso. 37
- Figura 5 – Imagem ilustrativa de uma região de correlação espacial para um correlograma pertencente a família de Wendland, definida a partir da Equação (2.2) com suavização $\kappa = 3$ 40
- Figura 6 – Gráfico ilustrativo do semivariograma, $\gamma(\mathbf{h})$ versus $\mathbf{h}$, para família Wendland com parâmetro de suavização $\kappa = 1$ 48
- Figura 7 – Gráficos de influência local para os dados simulados e perturbado. Em (a) é indicado o gráfico da curvatura normal, C_i para os dados simulados, enquanto em (b) a representação é para os dados após a perturbação versus ordem das observações. 62
- Figura 8 – Semivariograma experimental, 9(a), boxplot em 9(b) para variável produtividade da soja em ($t \text{ } ha^{-1}$) referente ao ano agrícola 2016/2017. 64

Figura 9 – <i>Post-plot</i> da distribuição espacial dos pontos em 10(a) e gráfico de envelope das funções semivariâncias em 10(b) para variável produtividade da soja em ($t \text{ ha}^{-1}$) referente ao ano agrícola 2016/2017.	65
Figura 10 – Influência local a partir da curvatura normal total $C_i \times$ ordem das observações (11(a)), e direção máxima da curvatura normal $ l_{max} \times$ ordem das observações (11(b)) para o modelo selecionado de produtividade da soja.	68
Figura 11 – Mapas para os valores previstos de produtividade da soja em ($t \text{ ha}^{-1}$) na safra 2016/2017 para a função de covariância de Wendland com suavização $\kappa = 3$, 12(a) com todos os pontos e 12(b) sem as observações #52, #61 e #69.	72
Figura 12 – Mapas para os valores previstos de produtividade da soja em ($t \text{ ha}^{-1}$) na safra 2016/2017 para a função de covariância de Wendland com suavização $\kappa = 3$, 13(a) sem a observação #52, e 13(b) sem a observação #61.	72
Figura 13 – Mapas para os valores previstos de produtividade da soja em ($t \text{ ha}^{-1}$) na safra 2016/2017 para a função de covariância de Wendland com suavização $\kappa = 3$ sem a observação #69.	73
Figura 14 – Grade de pontos com tamanhos 10×10 e 20×20 , Figuras 15(a) e 15(b), respectivamente, distribuídos com espaçamento de 1 (uma) unidade entre os pontos.	90
Figura 15 – Grade de pontos de um conjunto de dados reais sobre precipitação pluviométrica no primeiro semestre de 2021 no estado de Pernambuco, Brasil, referente a 157 estações e medição.	90
Figura 16 – Mapa da localização espacial de 154 estações meteorológicas no Estado de Pernambuco.	103
Figura 17 – Histograma de densidade e <i>BoxPlot</i> da variável precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil, Figuras 18(a) e 18(b), respectivamente.	105
Figura 18 – <i>Postplot</i> e envelope da distribuição espacial da variável precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil, Figuras 19(a) e 19(b), respectivamente.	105

Figura 19 – Gráfico de probabilidade meio-normal com envelope simulado (95% de confiança) do modelo de regressão, com distribuição marginal Gama, ajustado com estrutura de correlação Wendland com $\kappa = 3$, para variável resposta precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil (20(a)). Gráfico de influência local segundo uma perturbação na matriz de correlação de trabalho (20(b)).	108
Figura 20 – Histograma de frequências e <i>BoxPlot</i> da variável precipitação pluviométrica acumulada (<i>mm</i>) o mês de agosto de 2021 no estado de Pernambuco, Brasil, Figuras 21(a) e 21(b), respectivamente.	131
Figura 21 – Gráfico <i>post-plot</i> segundo o quartil dos dados (Figura 22(a)) e semivariograma empírico da variável precipitação pluviométrica acumulada (Figura 22(b)).	131
Figura 22 – Gráfico de dispersão e correlação linear de pearson das variáveis envolvidas no estudo.	132
Figura 23 – Resíduos para o modelo dos componentes discreto e contínuo, Figuras 24(a) e 24(b), respectivamente, contra seus respectivos valores ajustados para os dados de precipitação pluviométrica acumulada.	134
Figura 24 – Resíduos para o modelo completo, Figura 25(a), contra os valores ajustados. Gráfico normal de probabilidade com envelopes simulados para o modelo ajustado, Figura 25(b), para os dados de precipitação pluviométrica acumulada.	135
Figura 25 – Gráfico de índices de $ l_{max} $ (autovetor associado ao maior autovalor relacionado a maior curvatura normal) para ponderação de casos, Figura 26(a), e perturbação nas respostas, Figura 26(b), para os dados de precipitação pluviométrica acumulada.	136

LISTA DE TABELAS

Tabela 1 – Funções de correlação de Wendland generalizada quando se aumenta o parâmetro de suavização κ com as respectivas equivalências para o parâmetro ν da família Matérn.	31
Tabela 2 – Fator do Modelo para Wendland e Matérn.	44
Tabela 3 – Escala de classificação segundo a dependência espacial entre as amostras para família Wendland.	44
Tabela 4 – Escala de classificação segundo a dependência espacial entre as amostras para família Matérn.	45
Tabela 5 – Parâmetros estimados por ML para os dados de uma amostra selecionada dentre as 1000 simuladas, considerando a família de covariância Wendland generalizada com parâmetro de suavização $\kappa = 3$ e parâmetros originais $(\tau_1, \tau_2)^\top = (0, 15; 0, 3)^\top$, $\phi_3 = 1$ e $\beta_0 = 1$	62
Tabela 6 – Estatísticas descritivas para os dados da produtividade da soja (PROD) $[t \text{ ha}^{-1}]$ ano 2016/2017 de uma área comercial de 127,18 ha, de produção de grãos, localizada próximo ao município de Cascavel, região Oeste do Paraná (PR)	64
Tabela 7 – Parâmetros estimados por método máxima verossimilhança para modelo com covariância espacial de Wendland (W) e Matérn (M) para vários valores do parâmetro de suavização κ e ν , para os dados da variável produtividade da soja em $(t \times \text{ha}^{-1})$ do ano agrícola 2016/2017, com $\phi_3 = 1, 2$	67
Tabela 8 – Estatísticas descritivas para variável produtividade da soja (Prod($t \text{ ha}^{-1}$)) com e sem a presença das observações potencialmente influentes (Prod*($t \text{ ha}^{-1}$)), e quando excluídas individualmente #52, #61 e #69.	69
Tabela 9 – Parâmetros estimados por ML sem a presença individual e/ou conjunta de observações potencialmente influentes. Desvios padrões são apresentados entre parênteses.	70
Tabela 10 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Normal. Os pontos são distribuídos em grades regulares de tamanho 10×10 e 20×20 . Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.	93

Tabela 11 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Poisson. Os pontos são distribuídos em grades regulares de tamanho 10×10 e 20×20 . Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.	94
Tabela 12 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Gama. Os pontos são distribuídos em grades regulares de tamanho 10×10 e 20×20 . Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.	95
Tabela 13 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Normal. Os pontos são distribuídos em uma grade de um conjunto de dados reais sobre precipitação no estado de Pernambuco, Brasil. Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.	96
Tabela 14 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Poisson. Os pontos são distribuídos em uma grade de um conjunto de dados reais sobre precipitação no estado de Pernambuco, Brasil. Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.	97
Tabela 15 – Proporções (%) de certos na seleção da estrutura de correlação espacial para respostas correlacionada da Gama. Os pontos são distribuídos em uma grade de um conjunto de dados reais sobre precipitação no estado de Pernambuco, Brasil. Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.	98
Tabela 16 – Funções de ligação e derivadas da média (μ) utilizadas na expressão dos elementos da matriz Hessiana das medidas de influência local para várias distribuições da família exponencial.	100
Tabela 17 – Resultados para os esquemas de perturbação	102
Tabela 18 – Estatísticas descritivas para respostas médias de precipitação entre os meses de Janeiro e Junho de 2021 no estado de Pernambuco, Brasil.	104
Tabela 19 – Estimativas dos parâmetros do modelo geoestatístico para famílias Wendland, Matérn e modelo Esférico, os quais definem a matriz de correlação de trabalho, para variável resposta precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil.	107

Tabela 20 – Estimativas e erros padrão dos parâmetros do modelo de regressão $\eta_i = \beta_0 + \beta_1 x(\mathbf{s}_i) + \beta_2 y(\mathbf{s}_j) + \beta_3 x(\mathbf{s}_i)H(\mathbf{s}_i) + \beta_4 x^2(\mathbf{s}_i)$, ajustado com estrutura de correlação Wendland com $\kappa = 3$, para variável resposta precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil.	108
Tabela 21 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Wendland ($\kappa = 2$) e seus respectivos $\sqrt{EQM}$ e viés.	127
Tabela 22 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Matérn ($\nu = 1, 5$) e seus respectivos $\sqrt{EQM}$ e viés.	127
Tabela 23 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação Wendland ($\kappa = 2$) e seus respectivos $\sqrt{EQM}$ e viés.	127
Tabela 24 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Matérn ($\nu = 1, 5$) e seus respectivos $\sqrt{EQM}$ e viés.	128
Tabela 25 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Wendland ($\kappa = 2$) e seus respectivos $\sqrt{EQM}$ e viés.	128
Tabela 26 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Matérn ($\nu = 1, 5$) e seus respectivos $\sqrt{EQM}$ e viés.	128

Tabela 27 – Percentual de vezes em que a observação perturbada (caracterizada como influente) foi detectada sobre as 1000 réplicas, sob os esquema de perturbação ponderação de casos, variável resposta e covariáveis. A matriz de covariância que gerou os dados foi a Wendland ($\kappa = 2$). Considerou-se a medida de influência local total, correspondente a direção da i -ésima observação ($C_i = 2 b_{ii} $).	129
Tabela 28 – Percentual de vezes em que a observação perturbada (caracterizada como influente) foi detectada sobre as 1000 réplicas, sob os esquema de perturbação ponderação de casos, variável resposta e covariáveis. A matriz de covariância que gerou os dados foi a Matérn ($\nu = 1,5$). Considerou-se a medida de influência local total, correspondente a direção da i -ésima observação ($C_i = 2 b_{ii} $).	129
Tabela 29 – Medidas resumo para variável precipitação pluviométrica acumulada (Prec).	130
Tabela 30 – Estimativas dos parâmetros de regressão para o modelo quasi-gama aumentado com zero e espacialmente correlacionado pelo algoritmo <i>Expectation-Solution</i> (ES).	134
Tabela 31 – Estimativas e desvio percentual relativo para os parâmetros dos submodelos quando se é retirada observações potencialmente influentes. Erros padrão encontram-se entre parênteses.	137

LISTA DE ABREVIATURAS E SIGLAS

ACE	Área de continuidade (ou correlação) espacial
CIC	Crítério de informação de correlação
CNBC	Condition-Number-Based Criterion
CV	Coeficiente de variação
DPR	Desvio percentual relativo
DQMR	Desvio quadrático médio relativo
EEG's	Equações de estimação generalizadas
FM	Fator do modelo
GLM	Modelo linear generalizado
GW	Generalizada de Wendland
IDE	Índice de dependência espacial
MD	Máxima distância
MGLM	Modelos lineares generalizados mistos
ML	Máxima verossimilhança
OLS	Mínimos quadrados ordinários
QIC	Crítério quasi-verossimilhança sob o modelo de independência
RJC	Rotnizky and Jewell's criterion

SUMÁRIO

1	INTRODUÇÃO	19
1.1	OBJETIVOS E CONTRIBUIÇÕES DA TESE	21
1.1.1	Contribuições originais	21
2	INFERÊNCIA E ESTUDO DE DIAGNÓSTICOS DE INFLUÊNCIA EM MODELOS ESPACIAIS LINEARES GAUSSIANOS COM ESTRUTURA DE COVARIÂNCIA WENDLAND	23
2.1	RESUMO	23
2.2	INTRODUÇÃO	23
2.3	FAMÍLIA DE COVARIÂNCIA WENDLAND	27
2.3.1	Modelo teórico	32
2.3.2	Alcance prático e Área de Continuidade Espacial para as família de covariância Wendland generalizada e Matérn	38
2.3.3	Área de Correlação Espacial e Fator do Modelo	39
2.4	ESTIMAÇÃO DOS PARÂMETROS	45
2.4.1	Método dos Mínimos Quadrados Ordinários	45
2.4.2	Método de Máxima Verossimilhança	50
2.4.3	Erro padrão assintótico	52
2.5	DIAGNÓSTICO DE INFLUÊNCIA	54
2.5.1	Influência Local	55
2.5.2	Níveis de referência para classificar observações potencialmente influentes	60
2.6	SIMULAÇÃO	61
2.7	APLICAÇÃO	63
2.7.1	Modelagem geoestatística	65
2.7.2	Influência local	68
2.7.2.1	<i>Influência das observações potencialmente influentes nas estimativas dos parâmetros</i>	<i>69</i>
2.7.2.2	<i>Influência das observações potencialmente influentes na confecção dos Mapas de predição</i>	<i>70</i>
2.8	CONSIDERAÇÕES FINAIS DO CAPÍTULO	73

3	UM NOVO CRITÉRIO PARA SELEÇÃO DA MATRIZ DE CORRELAÇÃO DE TRABALHO EM MODELOS ESPACIAIS DE QUASI - VEROSSIMILHANÇA	74
3.1	RESUMO	74
3.2	INTRODUÇÃO	74
3.3	MODELO DE QUASI - VEROSSIMILHANÇA CORRELACIONADO	80
3.3.1	Estimação dos parâmetros de correlação espacial	82
3.3.2	Critérios para seleção da matriz de correlação	84
3.4	CONDITION-NUMBER-BASED CRITERION: CRITÉRIO PARA SELEÇÃO DA MATRIZ DE CORRELAÇÃO DE TRABALHO	85
3.4.1	Validação do critério de seleção de correlação espacial por estudo de simulação	88
3.5	MEDIDAS DE INFLUÊNCIA LOCAL	99
3.5.1	Esquemas de perturbação	100
3.5.1.1	<i>Perturbação na variável resposta e no preditor linear</i>	101
3.5.1.2	<i>Perturbação na matriz de correlação de trabalho</i>	101
3.6	APLICAÇÃO: ANÁLISE ESPACIAL DA PRECIPITAÇÃO PLUVIOMÉTRICA (MM) PARA O PRIMEIRO SEMESTRE DE 2021 NO ESTADO DE PERNAMBUCO	102
3.7	CONSIDERAÇÕES FINAIS DO CAPÍTULO	108
4	TÉCNICAS DE DIAGNÓSTICO EM MODELOS QUASI-GAMA AUMENTADA COM ZERO PARA DADOS ESPACIALMENTE CORRELACIONADOS	110
4.1	RESUMO	110
4.2	INTRODUÇÃO	110
4.3	MODELO QUASI-GAMA ESPACIALMENTE CORRELACIONADO	112
4.3.1	Estimação dos parâmetros	116
4.3.2	Erros padrão	117
4.4	TÉCNICAS DE DIAGNÓSTICO	118
4.4.1	Resíduos	119
4.4.2	Medidas de Influência local	120
4.4.2.1	<i>Esquema de perturbação ponderação de casos</i>	122

4.4.2.2	<i>Esquema de perturbação da variável resposta do componente contínuo do modelo Quasi-gama aumentada com zero e espacialmente correlacionado</i> .	122
4.4.2.3	<i>Esquema de perturbação simultânea nas variáveis preditoras do excesso de zeros e componente contínua</i>	123
4.5	ESTUDO DE SIMULAÇÃO	125
4.6	APLICAÇÃO: CHUVA ACUMULADA (MM) DO MÊS DE AGOSTO DE 2021 NO ESTADO DE PERNAMBUCO (BR)	129
4.7	CONSIDERAÇÕES FINAIS DO CAPÍTULO	138
5	CONCLUSÕES GERAIS	139
5.1	TRABALHOS FUTUROS	140
	REFERÊNCIAS	142
	APÊNDICE A – RESULTADOS DE ML	151
	APÊNDICE B – SISTEMAS DE EQUAÇÕES PARA O MÉTODO DOS MÍNIMOS QUADRADOS	153
	APÊNDICE C – DERIVADAS DA FUNÇÃO DE COVARIÂNCIA DE WENDLAND	155
	APÊNDICE D – DETALHES DOS RESULTADOS APRESENTADOS NA SEÇÃO 3.5 PARA MEDIDAS DE INFLUÊNCIA LOCAL	157
	APÊNDICE E – DERIVADAS PARA FUNÇÕES DE COVARIÂNCIA DE WENDLAND, COM RELAÇÃO A PERTURBAÇÃO ω_i, PARA VÁRIOS VALORES DE κ E μ	159

1 INTRODUÇÃO

O presente trabalho de tese está relacionado ao estudo de fenômenos que apresentam dependência espacial. Para tanto, nos concentramos nas ferramentas da Geoestatística, uma área da Estatística Espacial fundamentada nas ideias do matemático e geólogo Matheron (1965) quando estabeleceu a teoria das variáveis regionalizadas, construídas a partir de técnicas de estimação estatística, com aplicações em estudos cujo atributo de interesse encontra-se espacialmente distribuído. Com esta abordagem, desde os primeiros trabalhos nos estudos de mineração, a análise geoestatística vem sendo amplamente utilizada por profissionais das ciências agrárias, biológicas, geologia, engenharias, dentre outras áreas do conhecimento.

A Geoestatística apresenta como principal característica reunir os conceitos de variáveis aleatórias à variáveis espacialmente distribuídas, considerando sua localização geográfica e a dependência espacial, pressupondo que observações com localizações próximas tendem a serem mais semelhantes que as mais distantes. Portanto, faz sentido induzir que a distribuição das diferenças de variáveis amostradas em duas localizações distintas é constante em toda área, dependendo somente da distância entre elas e eventualmente da orientação dos pontos. Nesse sentido, a distribuição espacial de uma variável de interesse z em uma região contínua S no espaço d -dimensional fica caracterizada por uma realização de um processo estocástico $\{Z(s) : s \in S \subset \mathbb{R}^d\}$, com $d = 1, 2$ ou 3 . Portanto, o enfoque principal é fazer predições para os valores de $Z(s)$ em determinada localização espacial $s \in \mathbb{R}^d$, a partir de informações de seus vizinhos. Comumente considera-se $d = 2$ descrevendo *latitude* e *longitude* ou $d = 3$ para *latitude*, *longitude* e *altitude*. Particularmente, consideramos $d = 2$.

Dessa maneira, a predição de determinado valor de interesse está condicionada a localização dos dados amostrados, e a correlação espacial existente entre eles. Portanto, fica estabelecida uma função de correlação espacial, que possibilita explicar e/ou quantificar a variabilidade espacial dos atributos de interesse em uma região contínua $S \in \mathbb{R}^d$. Sendo assim, a partir da função de correlação espacial, com condições satisfeitas de estacionariedade de segunda ordem, é possível definir as funções de covariância e semivariância espaciais. Vale destacar que, do ponto de vista descritivo as três funções fornecem a mesma informação com relação a dependência espacial, embora a semivariância apresente-se de forma invertida. Ademais, é permitido a modelagem da parte não estocástica do fenômeno a partir da média, caracterizando o efeito da tendência.

Em suma, este trabalho encontra-se dividido em cinco Seções. Contudo, destacamos que, embora as Seções estejam relacionados pelo fato de estarem sobre o enfoque da análise e inferência em modelos geoestatísticos, cada Seção poderá ser lida de forma independente, inclusive com relação as notações conceituais. Na Seção 2 realizamos inferência em modelos espaciais lineares utilizando uma estrutura de covariância da família Wendland. Embora, esta abordagem com família Wendland não seja totalmente inédita, Wu (1995), Wendland (1995), Gneiting (2002a), Bevilacqua et al. (2019), pouco se foi explorado sobre sua aplicabilidade em análise de dados espaciais. Diante disso, são revistas suas propriedades estudando o comportamento do decaimento de sua função de correlação para vários valores do parâmetro de suporte. Desenvolvemos expressões para o alcance prático da família Wendland para vários valores de suavização e Matérn com suavização $\nu = 3, 5$. Estendemos as expressões para o Índice de Dependência Espacial (Índice de dependência espacial (IDE)) para famílias Wendland e Matérn. Ainda, desenvolvemos inferência baseada nos estimadores de máxima verossimilhança. Adicionalmente, avaliamos a sensibilidade dos estimadores de máxima verossimilhança para pequenas perturbações, em que são desenvolvidas medidas de influência local, e propomos caracterizar observações como potencialmente influentes utilizando o algoritmo *Jackknife-after-Bootstrap*.

Na Seção 3, consideramos modelos de quasi-verossimilhança sob Equações de Estimação Generalizadas (Equações de estimação generalizadas (EEG's)), e propomos um novo critério de seleção da matriz de correlação de trabalho, baseado no condicionamento da matriz de variância-covariância *naive*. A avaliação da performance do método é realizada por um extenso estudo de simulação, utilizando as distribuições marginais da Normal, Poisson e Gama para dados espacialmente correlacionados. A especificação da estrutura de correlação é baseada em funções de semivariâncias, utilizando as famílias Wendland, Matérn e Esférica.

Na Seção 4, utilizamos uma mistura da distribuição Bernoulli degenerada em zero com a Gama, proposto por Feuerverger (1979), então, à estendemos para derivar um modelo para dados espacialmente referenciados utilizando a ideia de quasi-verossimilhança. Com isso, derivamos equações de estimação para estimar os parâmetros do modelo de regressão da parte contínua, dada pela quasi-verossimilhança da distribuição gama, e os parâmetros que determinam a probabilidade de ocorrência de um valor zero nos dados observados (parte discreta). Adicionalmente, desenvolvemos medidas de diagnóstico para os resíduos e de influência local para o modelo desenvolvido. Ainda, como contribuição de aplicação utilizamos um conjunto de dados para validar a aplicabilidade do modelo em explicar a precipitação pluviométrica acumulada para o mês de agosto de 2021 no estado de Pernambuco, Brasil. Finalmente, na

Seção 5 apresentamos as considerações finais e propostas de trabalhos futuros.

Destarte, como ferramenta computacional para obter os resultados numéricos, análise de diagnóstico de influência local e interpolação espacial utiliza-se o *Software R Core Team (2020). R: A language and environment for statistical computing*, (R Core Team, 2020), com funções criadas pelo autor e eventualmente recorrendo ao uso de pacotes disponíveis no *Software*, estes, quando utilizados são citados ao longo do texto.

1.1 OBJETIVOS E CONTRIBUIÇÕES DA TESE

O objetivo principal deste trabalho de tese é realizar uma releitura das principais ferramentas metodológicas para análise e inferência em dados georreferenciados. Com isso, dispomos das seguintes metas específicas:

1. Apresentar a aplicabilidade das funções da família Wendland, uma classe de funções de covariância compactamente suportada, em comparação com a família Matérn, para modelagem de modelos espaciais lineares Gaussianos;
2. Apresentar uma escala de classificação segundo a dependência espacial e considerando o IDE entre as amostras para as famílias Wendland e Matérn;
3. Estender a metodologia dos modelos de quasi-verossimilhança para dados espaciais considerando distribuições da família exponencial;
4. Desenvolver um novo critério para seleção da matriz de correlação de trabalho para modelos espaciais de quasi-verossimilhança sob EEG;
5. Propor um modelo geoestatístico para análise de dados contínuos com assimetria positiva e excesso de zeros;
6. Avaliar a metodologia desenvolvida com aplicações a conjuntos de dados atuais.

1.1.1 Contribuições originais

Este trabalho de tese incorpora as seguintes contribuições originais:

1. Expressões para o alcance prático da família Wendland para vários valores de suavização e família Matérn com suavização $\nu = 3,5$. Índice de Dependência Espacial (IDE)

para famílias Wendland e Matérn. Inferência e diagnóstico em modelos espaciais lineares Gaussianos com a Wendland em comparação com Matérn. Classificação de observações como potencialmente influentes, sob medidas de influência local, a partir de uma generalização do algoritmo *Jackknife-after-Bootstrap*;

2. Desenvolvimento de um novo critério para seleção da matriz de correlação de trabalho para modelos espaciais de quasi-verossimilhança sob EEG, a partir da norma matricial e número de condição das matrizes de variâncias-covariâncias;
3. Desenvolvimento de expressões para medidas de influência local para modelos espaciais de quasi-verossimilhança sob os esquemas de perturbação no preditor linear e na matriz de correlação de trabalho, considerando as distribuições Binomial, Poisson, Binomial negativa, Gama e Normal inversa;
4. Desenvolvimento de um modelo geoestatístico quasi-gama para modelar dados com assimetria positiva e excesso de zeros;
5. Estimação dos parâmetros para o modelo geoestatístico quasi-gama utilizando o algoritmo Expectation-Solution (ES). Desenvolvimento de expressões para os erros-padrão, medidas para os resíduos e diagnóstico de influência local.

2 INFERÊNCIA E ESTUDO DE DIAGNÓSTICOS DE INFLUÊNCIA EM MODELOS ESPACIAIS LINEARES GAUSSIANOS COM ESTRUTURA DE COVARIÂNCIA WENDLAND

2.1 RESUMO

Nesta Seção, nos concentramos nos modelos lineares espaciais Gaussianos comparando a família de covariância Wendland com a família Matérn. Calculamos um índice de dependência espacial para ambas as famílias de covariância e estabelecemos relações entre o alcance prático e o parâmetro que controla o decaimento de suas funções de correlação. Uma abordagem de diagnóstico de influência local é usada para avaliar a sensibilidade dos estimadores de máxima verossimilhança. Para generalizar o critério de validação cruzada usual, propomos usar o critério *Jackknife-após-Bootstrap* para obter os níveis de referência ou *cutoff* estabelecidos para determinar se uma observação é potencialmente influente ou não. Uma aplicação a um conjunto de dados sobre o rendimento da soja ilustra a metodologia desenvolvida aqui. Finalmente, concluímos que os resultados para a família de covariâncias de Wendland são satisfatórios sob os pressupostos do modelo linear espacial Gaussiano.

2.2 INTRODUÇÃO

Em diversas áreas do conhecimento a exemplo de Ciências agrárias, biológicas, sociais, Engenharias (ambiental, civil, sanitária, etc), Epidemiologia, Meteorologia, Geologia e Mineração, agrimensura, zootecnia, dentre tantas outras, estão presentes fenômenos com dados correlacionados e georreferenciados, em que é possível a replicação de dados contínuos ou não em um mapa de distribuição espacial de continuidade. Alguns dos exemplos são mapas da variabilidade espacial da produtividade de determinada cultura (soja, milho, café, etc), volume de um minério, poluição do ar em uma cidade ou região, caracterização de áreas de riscos a saúde pública, entre outros. Neste contexto, surge o conceito de interpolação espacial à qual pressupõe que os valores dos atributos de interesse assemelham-se com a proximidade entre si.

Com isso, é necessário introduzir o conceito de função de covariância espacial, a qual fornece informações importantes sobre a estrutura de dependência espacial do processo aleatório subjacente, sendo responsável por determinar a performance da previsão espacial. Portanto,

é fundamental que sua escolha seja realizada com cautela de modo a produzir resultados fidedignos no que concerne a realização da variável aleatória em diferentes locais da região em estudo.

Estão disponíveis na literatura uma diversidade de técnicas que realizam a interpolação espacial. No entanto, no âmbito da Geoestatística, o qual este trabalho de tese está inserido, a interpolação de um processo aleatório espacialmente correlacionado é realizada pelo melhor preditor linear não viesado (*best linear unbiased estimator*), comumente nomeado de Krigagem. Esta técnica é empregada através da variação amostral associada a distância entre as observações, baseando-se na matriz de covariância espacial indexada por um vetor de parâmetros desconhecidos e, sendo necessário a solução de um sistema, geralmente grande, de equações lineares. Desta forma, o ajuste do modelo depende da inversão da matriz de covariância, o que torna o processo custoso computacionalmente, sobretudo quando o número de observações é grande.

São várias as abordagens disponíveis para descrição de modelos de covariâncias espaciais válidas (definidas positivas), a partir dos quais é permitido a formulação de modelos paramétricos de semivariâncias. Dentre estes, os mais populares são os modelos Esférico, Exponencial, Gaussiano e Matérn. Embora as funções do tipo Matérn tenham sido introduzidas na década de 60 por Bertil Matérn (1960), no contexto da análise de dados espaciais ganharam notoriedade a partir dos trabalhos de Stein (1999) e Diggle e Ribeiro (2000), sendo então nomeadas de família Matérn.

A suavização do processo subjacente com estrutura de covariância da família Matérn é determinada por um parâmetro $\nu > 0$, específico desta família, nomeado de parâmetro de suavização, ver Diggle e Ribeiro (2000). Segundo os autores anteriormente citados, este parâmetro controla a diferenciabilidade do processo. Assim, desde sua investidura na modelagem de dados espaciais, esta família vem sendo amplamente utilizada como melhor função de correlação espacial, tendo em vista sua flexibilidade. Apesar disso, do ponto de vista computacional esta família de covariância apresenta limitações quando se trabalha com grandes conjuntos de dados, haja vista a geração de matrizes densas, pois apresentam suporte global, melhor dizendo, suas correlações são não-nulas em todo domínio de definição.

À vista disso, uma alternativa interessante para melhoria da eficiência computacional é introduzir funções de covariâncias espaciais que apresentem um suporte compacto, isto é, as correlações assumem valores nulos a partir de uma certa distância. Desta maneira, são obtidas matrizes de covariância esparsas, o que é benéfico para redução do custo computacional, ver

Wendland (1995), Buhmann (2003), Wendland (2004), Fasshauer (2007), Moreaux (2008), Porcu et al. (2013). Contudo, são poucos os modelos de funções correlações compactamente suportadas disponíveis na literatura, sobretudo quando utilizadas em aplicações de fenômenos espaciais. Diante disso, propomos neste trabalho estudar e realizar inferências em modelos espaciais, utilizando uma função de correlação compactamente suportada definida no artigo de Bevilacqua et al. (2019), nomeada de Generalizada de Wendland, em que a partir de um parâmetro de suavização ($\kappa > 0$) são amplamente flexíveis, colocando-a como uma competidora natural da família Matérn.

As funções de Wendland são procedentes da popular classe de funções de base radial – utilizadas em aplicações de aproximação numérica – são compactamente suportadas e definidas no intervalo $[0, 1)$. Introduzidas inicialmente por Wu (1995), e posteriormente por Wendland (1995), podendo ser definidas a partir da distância euclidiana, são facilmente estendidas para geração de funções de covariância radialmente simétricas as quais são importantes para inferência e predição de campos aleatórios. Da forma como são definidas, as funções de Wendland destacam-se por permitirem o aumento da eficiência computacional, além de reduzir o risco de mal condicionamento da matriz de covariância espacial (matriz de interpolação), visto que seus valores são nulos para distâncias maiores que um certo parâmetro preestabelecido, nomeado suporte, o que é essencial em análise geoestatística. Tal característica tem o potencial de gerar matrizes esparsas, facilitando a solução numérica do problema de interpolação, sobretudo no procedimento de Krigagem em que requer a inversão de matrizes, geralmente de alta dimensão.

A literatura é rica em trabalhos que empregam o uso de funções de base radial compactamente suportadas no estudo de aproximações em várias áreas da ciência, sobretudo em aplicações que exigem aproximações de equações diferenciais parciais por elementos finitos. No entanto, sua aplicabilidade para análise de dados espaciais é ainda recente com poucos trabalhos disponíveis. Neste contexto, alguns dos trabalhos mais recentes são: Gneiting (2002a) que propôs uma classe de funções de correlação compactamente suportada, incluindo a Wendland, que parametrizam a suavidade do processo aleatório estacionário e isotrópico associado. Gneiting (2002b) em que são apresentadas funções estacionárias e não separáveis para processos aleatórios espaço-temporais. Em Myers et al. (2002) são construídas funções de base radial espaço-tempo com demonstrações a partir de exemplos. Moreaux (2008) revisa as principais funções de covariância com suporte compacto. Porcu et al. (2013) fornecem uma metodologia para análise geoestatística multivariada com funções radiais de suporte compacto, baseados

nas funções de Askey e por *B-splines*. Chernih e Hubbert (2014) mostram as condições necessárias e suficientes para que as funções de Wendland sejam positivas definidas. No mesmo ano, Chernih, Sloan e Womersley (2014) mostram que as funções de Wendland convergem uniformemente para uma Gaussiana quando o parâmetro de suavização tende ao infinito. Daley, Porcu e Bevilacqua (2015) exploram funções de covariância com suporte compacto para campos aleatórios multivariados. Genton e Kleiber (2015) apresentam funções de covariância cruzada no estudo de geoestatística multivariada, incluindo suporte compacto e domínios esféricos. Liang e Marcotte (2016) descrevem uma classe de funções não estacionária com suporte compacto. Bevilacqua et al. (2019) estudam estimação e predição de campos aleatórios Gaussianos com modelos de covariância da família Wendland. Porcu, Bevilacqua e Genton (2020) descrevem funções de covariância espaço-temporais não separáveis utilizando a família Wendland. Bachoc et al. (2020) fornecem condições para equivalências de medidas associadas a campos aleatórios Gaussianos multivariados, com enfoque nas funções Matérn e Wendland. E Bevilacqua, Diggle e Porcu (2020) propõem uma classe de funções de covariância para campos aleatórios bivariados definidas em esferas.

Embora os autores anteriormente mencionados tenham fornecido várias abordagens metodológicas, sobre os vários caminhos possíveis de se estudar e/ou aplicar as funções radiais com suporte compacto, na modelagem de dados espaciais ainda carece de tratamento destas funções, sobretudo da família Wendland. Diante disso, neste trabalho aproveitamos a capacidade das funções de Wendland em gerar matrizes esparsas, bem como seu potencial de eficiência computacional para realizar um estudo de inferência e diagnóstico em modelos espaciais lineares Gaussianos. Então, nosso enfoque é avaliar a aplicabilidade da covariância de Wendland em predições espaciais, a partir de um exemplo numérico com dados de produtividade de soja em comparação com a família Matérn.

Outrossim, são revistas suas propriedades estudando o comportamento do decaimento de sua função de correlação para vários valores do parâmetro de suporte, que é utilizado para definir o raio de dependência espacial. Desenvolvemos expressões para o alcance prático da família Wendland para vários valores de suavização e Matérn com suavização $\nu = 3,5$. A partir das ideias de Seidel e Oliveira (2015) expandimos expressões para o Índice de Dependência Espacial (IDE) para famílias Wendland e Matérn. Então, realizamos o processo de inferência e diagnóstico em modelos espaciais Gaussianos com a Wendland como matriz de covariância. Assim, com objetivo de avaliar a sensibilidade dos estimadores de máxima verossimilhança para pequenas perturbações, desenvolvemos medidas de influência local, e propomos

caracterizar observações como potencialmente influentes a partir de uma generalização do algoritmo *Jackknife-after-Bootstrap* definido em Martin e Roberts (2010). Para finalizar o Capítulo, realiza-se um pequeno estudo de simulação e, um exemplo de aplicação a dados reais relacionados a agricultura de precisão é empregado.

2.3 FAMÍLIA DE COVARIÂNCIA WENDLAND

Considere um vetor ou campo aleatório multivariado d –dimensional $\mathbf{Z} = (Z_1(\mathbf{s}_1), \dots, Z_n(\mathbf{s}_n))$, definido em uma região/localização espacial $\mathbf{S} \subset \mathbb{R}^d, d \geq 1$, particularmente $d = 2$, em que $Z_i(\mathbf{s}_i)$ denota a i –ésima variável coletada na localização espacial $\mathbf{s}_i \in \mathbf{S}$, com $i = 1, \dots, n$. Então, com a suposição de que $\mathbf{Z}$ segue uma distribuição Gaussiana com vetor de média nulo e estacionário de segunda ordem, tem-se que $\mathbf{Z}$ é completamente caracterizado por sua estrutura de covariância cruzada $\mathbf{C} \in \mathbb{R}^{n \times n}$, isto é:

$$\mathbf{C}(\mathbf{s}, (\mathbf{s} + \mathbf{h})) = Cov(Z_i(\mathbf{s}), Z_j(\mathbf{s} + \mathbf{h})) = C_{ij}(\mathbf{s} - (\mathbf{s} + \mathbf{h})), \quad (2.1)$$

ou seja, a covariância entre duas observações medidas em duas localizações espaciais $\mathbf{s}$ e $\mathbf{s} + \mathbf{h}$, depende apenas do vetor de distâncias que os separa $\mathbf{h}$.

A função de covariância definida em (2.1) caracteriza-se por mensurar a dependência espacial entre as observações. No entanto, sua definição nem sempre é uma tarefa fácil, tendo em vista a necessidade de satisfazer a propriedade de ser positiva definida. Diante disso, aqui considera-se sua construção a partir das funções de correlação do tipo Generalizadas de Wendland (Generalizada de Wendland (GW)) definidas em Bevilacqua et al. (2019) como na Equação (2.2).

$$\varphi_{\mu, \kappa}(r) = \begin{cases} \frac{1}{B(2\kappa, \mu+1)} \int_r^1 x(x^2 - r^2)^{\kappa-1} (1-x)^\mu dx & \text{se } 0 \leq r < 1 \\ 0 & \text{se } r \geq 1, \end{cases} \quad (2.2)$$

em que $B(\cdot)$ denota a função Beta, e $r = \|\mathbf{h}\|$ denota a distância euclidiana entre duas localizações espaciais $\mathbf{s}$ e $\mathbf{s} + \mathbf{h}$, com $\|\cdot\|$ a norma euclidiana no $\mathbb{R}^2$. Os parâmetros μ e κ definem a positividade e suavização da função em um espaço d –dimensional. Funções deste tipo tem a vantagem de suas transformadas de Fourier serem positivas, o que garante a definição de positividade. Este fato decorre do Teorema de *Bochner* que pode ser consultado em Fasshauer (2007). Apesar disto, em Gneiting (2002a), pode-se verificar que para $\kappa \in \{0, 1, 2, \dots\}$, $\varphi_{\mu, \kappa}$,

é definida positiva se, e somente se satisfaz a condição (2.3).

$$\mu \geq \frac{d+1}{2} + \kappa. \quad (2.3)$$

As funções de GW caracterizam por seus valores diminuírem ou aumentarem, monotonicamente, em relação a um certo parâmetro $\beta > 0$, previamente arbitrado denominado de suporte. Portanto, este parâmetro está relacionado ao preenchimento da matriz de covariância. Quando seu valor é muito pequeno o condicionamento da matriz melhora, por outro lado, a qualidade de replicação dos pontos amostrados é prejudicada. No entanto, a medida que seu valor aumenta, obtém-se um melhor comportamento de replicação, mas a condição da matriz aumenta implicando no custo computacional. Portanto, uma escolha razoável para o suporte é aquela que representa um meio termo entre estabilidade e boa qualidade de aproximação, a qual veremos ao longo deste Capítulo.

Posto isto, de agora em diante o suporte é dimensionado substituindo o raio r por h/β , com $h = \|\mathbf{h}\|$. Assim, assume que a covariância associada em (2.1), com suporte compacto $\beta > 0$, variância σ^2 e parâmetro de suavidade $\kappa > 0$ pode ser escrita como:

$$\mathbf{C}(h) = \sigma^2 \varphi_{\mu, \kappa}(h/\beta), \quad h \geq 0. \quad (2.4)$$

Claramente o cálculo de (2.4) utilizando (2.2) é preciso recorrer a métodos/técnicas de integração numérica. Contudo, como aponta Bevilacqua et al. (2019), torna-se inviável tais métodos no sentido de que trabalha-se com modelos de covariância paramétrica de forma fechada. Em contrapartida, obtém-se uma solução de forma analítica para família de funções definida em 2.2 quando o parâmetro de suavidade κ assumir um valor inteiro positivo.

De maneira mais clara, as construções de funções de correlação pertencentes a família (2.2), dependem de transformações entre os elementos da classe funções contínuas positivas que preservam o suporte compacto. Estas transformações, apresentada em Gneiting (2002a), sugere o uso de dois operadores, anteriormente já definidos em Matheron (1965), denominados operador integral (*montée*) e operador diferencial inverso (*descente*).

Na literatura são apresentadas algumas das funções de Wendland para vários valores de κ utilizando estes operadores. No entanto, são omitidos os procedimentos de obtenção das mesmas. Dessa maneira, aqui desenvolvemos, a título de exemplo, os cálculos envolvidos para se obter, de forma explícita, as funções de correlação da família GW a partir do procedimento de integração por partes. Sem perda de generalidade, em todos os cálculos considera-se $\beta = \sigma^2 = 1$ e parâmetro de suavidade $\kappa = 0$ e 1. Para $\kappa = 0$ o resultado segue direto, obtido

aplicando a função potência truncada $(1-r)_+^\mu = (1-r)^\mu$, se $0 \leq r < 1$. Para $\kappa = 1$ e utilizando (2.2), obtém-se,

$$\begin{aligned}
 \varphi_{\mu,1,1,1}(r) &= \frac{1}{B(2 * 1, \mu + 1)} \int_r^1 x(x^2 - r^2)^{1-1} (1-x)^\mu dx \\
 &= \frac{1}{B(2, \mu + 1)} \int_r^1 x(1-x)^\mu dx \\
 &\stackrel{(*)}{=} \frac{1}{\frac{\Gamma(2)\Gamma(\mu+1)}{\Gamma(2+\mu+1)}} \left[\frac{r}{\mu+1} (1-r)^{\mu+1} + \frac{1}{\mu+1} \frac{1}{\mu+2} (1-r)^{\mu+2} \right] \\
 &\stackrel{(**)}{=} \frac{1}{\frac{1}{(\mu+1)(\mu+2)}} \frac{1}{(\mu+1)(\mu+2)} (1-r)^{\mu+1} [r(\mu+1) + 1] \\
 &= (1-r)^{\mu+1} [r(\mu+1) + 1]. \tag{2.5}
 \end{aligned}$$

Na passagem $\stackrel{(*)}{=}$ utiliza-se as propriedades da função Beta e integração por partes, enquanto em $\stackrel{(**)}{=}$ propriedades da função Gama. Uma forma alternativa de obter estes resultados é aplicar o operador integral como segue:

$$\begin{aligned}
 \int_r^1 x(1-x)^\mu dx &= -\frac{x}{\mu+1} (1-x)^{\mu+1} \Big|_r^1 + \frac{1}{\mu+1} \int_r^1 (1-x)^{\mu+1} dx \\
 &= \frac{r}{\mu+1} (1-r)^{\mu+1} + \frac{1}{\mu+1} \left[-\frac{1}{\mu+2} (1-x)^{\mu+2} \right]_r^1 \\
 &= \frac{r}{\mu+1} (1-r)^{\mu+1} + \frac{1}{\mu+1} \frac{1}{\mu+2} (1-r)^{\mu+2} \\
 &= \frac{1}{\mu+1} (1-r)^{\mu+1} \left[r + \frac{1}{\mu+2} (1-r) \right] \\
 &= \frac{1}{\mu+1} \frac{1}{\mu+2} (1-r)^{\mu+1} [r(\mu+1) + 1],
 \end{aligned}$$

ainda,

$$\begin{aligned}
 \int_0^1 x(1-x)^\mu dx &= -\frac{x}{\mu+1} (1-x)^{\mu+1} \Big|_0^1 + \frac{1}{\mu+1} \int_0^1 (1-x)^{\mu+1} dx \\
 &= \frac{1}{\mu+1} \left[-\frac{1}{\mu+2} (1-x)^{\mu+2} \right]_0^1 \\
 &= \frac{1}{\mu+1} \frac{1}{\mu+2}.
 \end{aligned}$$

Portanto, obtemos

$$\begin{aligned}
 \varphi_{\mu,\kappa,1,1}(x) &= \frac{\frac{1}{\mu+1} \frac{1}{\mu+2} (1-r)^{\mu+1} [r(\mu+1) + 1]}{\frac{1}{\mu+1} \frac{1}{\mu+2}} \\
 &= (1-r)^{\mu+1} [r(\mu+1) + 1].
 \end{aligned}$$

Que é o mesmo resultado obtido em (2.5). Os demais casos são obtidos similarmente aplicando (2.2) ou mediante aplicação repetida do operador integral, os quais estão apresentadas na

Tabela 1. Ademais, segundo Gneiting (2002a), as funções pertencentes a família de Wendland são 2κ diferenciáveis na origem, positiva e estritamente decrescente em seu suporte compacto. Ainda, Gneiting (2002a) estabelece uma relação de equivalência no comportamento entre as funções de Wendland e Matérn em função dos respectivos parâmetros de suavidade, κ e ν , sendo $\nu = \kappa + \frac{1}{2}$. Bevilacqua et al. (2019) define a classe de funções pertencentes a família Matérn, equação (2.6), e faz ver as expressões de equivalência com os casos da Wendland, ver tabela 1.

$$\mathcal{M}_{\nu,\alpha,\sigma^2}(r) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r}{\alpha}\right)^\nu \mathcal{K}_\nu\left(\frac{r}{\alpha}\right), \quad r \geq 0, \quad (2.6)$$

em que ν é o parâmetro de suavização, σ^2 a variância e α o parâmetro de escala positivo. De forma análoga ao parâmetro de suavização para GW κ , o parâmetro ν para família Matérn caracteriza a diferenciabilidade na origem e, como consequência, a diferenciabilidade dos caminhos de amostra associados. Em particular para um número inteiro positivo κ , os caminhos de amostra são κ vezes diferenciáveis, em qualquer direção, se e somente se $\nu > \kappa$, (BEVILACQUA et al., 2019).

A classe de funções definida em (2.6) tem sido muito popular em estudos espaciais devido sua flexibilidade. Sua origem deu-se no trabalho do estatístico sueco Bertil Matérn, Matérn (1960), quando escreveu sua tese de doutorado na década de 60. Daí a origem do nome família Matérn. Apesar disso, no que diz respeito a aplicação em análise de dependência espacial, as funções desta família foram apresentadas por Diggle e Ribeiro (2000) como uma estrutura de correlação espacial através de uma modelagem utilizando a distribuição de probabilidade Gaussiana. Esta família contém, como casos particulares, a função de covariância Exponencial e Gaussiana quando $\nu = \frac{1}{2}$ e $\nu \rightarrow \infty$, respectivamente, (DIGGLE; RIBEIRO, 2000; GNEITING, 2002a). Similarmente, Chernih, Sloan e Womersley (2014) demonstram que se $\kappa \rightarrow \infty$, então, a partir de um redimensionamento adequado as funções de Wendland convergem uniformemente para uma Gaussiana.

Sendo assim, na Tabela 1 constam algumas das funções até a 3ª ordem da família Wendland, originadas a partir de Wendland (1995), e suas respectivas equivalências com a família Matérn. Outras caracterizações podem ser vistas em Wu (1995), Gneiting (2002a), Schaback (2011), Chernih e Hubbert (2014), Chernih, Sloan e Womersley (2014) e Porcu, Zastavnyi e Bevilacqua (2017).

Tabela 1 – Funções de correlação de Wendland generalizada quando se aumenta o parâmetro de suavização κ com as respectivas equivalências para o parâmetro ν da família Matérn.

κ	$\varphi_{\mu,\kappa,1,1}$	ν	$\mathcal{M}_{\nu,1,1}(r)$
0	$(1-r)_+^\mu$	0,5	e^{-r}
1	$(1-r)^{\mu+1} [r(\mu+1) + 1]$	1,5	$(1+r)e^{-r}$
2	$\left(1 + (\mu+2)r + \frac{1}{3}((\mu+2)^2 - 1)r^2\right) (1-r)^{\mu+2}$	2,5	$\left(1 + r + \frac{r^2}{3}\right)e^{-r}$
3	$(1-r)^{\mu+3} \left(1 + r(\mu+3) + r^2(2\mu^2 + 12\mu + 15)\frac{1}{5} + r^3(\mu^3 + 9\mu^2 + 23\mu + 15)\frac{1}{15}\right)$	3,5	$\left(1 + \frac{r}{2} + \frac{6r^2}{15} + \frac{r^3}{15}\right)e^{-r}$

Fonte: Elaborada pelo autor (2022).

Considera-se $\beta = \sigma^2 = 1$ e r denota a distância euclidiana entre duas localizações espaciais. O parâmetro μ caracteriza o comportamento da concavidade da curva, além de definir a positividade da matriz de correlação.

Note que na Tabela 1, por simplicidade, os resultados são para os casos particulares dos parâmetros de escala $\beta = 1$ e $\alpha = 1$, para as famílias de covariância GW e Matérn, respectivamente, pertencentes ao espaço d -dimensional, $\mathbb{R}^2(d = 2)$. Considera-se o parâmetro $\mu = \frac{d+1}{2} + \kappa + x$, com $x > 0$, assim, garante a positividade da matriz de covariância estabelecida pela relação (2.3), além disso, este parâmetro caracteriza o comportamento da concavidade da curva de correlação e semivariância para família GW. No entanto, uma relação de equivalência entre os parâmetros β e α foi estabelecida por Bevilacqua et al. (2019), como indica a equação (2.7), no qual dado um dos parâmetros é possível determinar o outro.

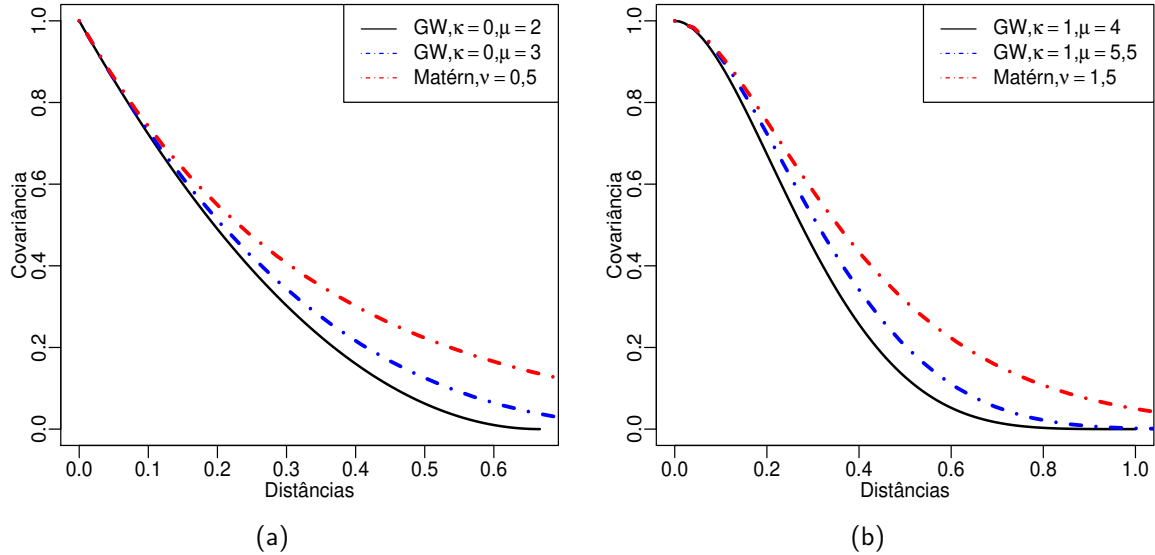
$$\beta^{2\kappa+1} = \frac{\mu\Gamma(2\kappa + \mu + 1)\sigma_1^2\sigma_2^{-2}\alpha^{2\nu}}{\Gamma(\mu + 1)}, \quad (2.7)$$

em que σ_1^2 e σ_2^2 denotam as variâncias das famílias GW e Matérn, respectivamente.

Diante disto, como ilustração do comportamento de equivalências entre a família GW e Matérn, Tabela 1, na Figura 1 tem-se os gráficos para os casos específicos de $\kappa = 0$ e $\kappa = 1$ com os valores de μ seguindo o critério para positividade da correlação GW. É considerado um alcance fixo igual a 1 para família Matérn. No caso da Wendland, como dito no início da seção, o alcance é controlado pelo parâmetro de escala, β . Na seção 2.3.1 é possível verificar a influência de β na amplitude da dependência espacial.

Uma particularidade entre as famílias de covariância GW e Matérn é que ambas permitem uma parametrização contínua da suavidade do campo aleatório gaussiano subjacente. Apesar disso, segundo Gneiting et al. (2013), Jeong e Jun (2015) as funções da classe Matérn não funcionam bem para grandes distâncias na casca esférica, sendo válidas apenas se, e somente se o parâmetro de suavidade satisfizer a condição $0 < \nu \leq 0,5$ para esfera uni, bidimensional ou tridimensional.

Figura 1 – Modelos de correlação compatíveis para o caso $d = 2$. Para o modelo Matérn os parâmetros de suavização $\nu = 0,5$ e $\nu = 1,5$ com equivalência para Wendland $\kappa = 0$ e $\kappa = 1$. Equivalências entre os parâmetros de escala em (a) com $\alpha = 1/3$, fixo, $\beta = 0,667$ e $1,00$; em (b) $\alpha = 1/4,75$, fixo, $\beta = 1,038$ e $1,356$, calculados pela Equação (2.7).



Fonte: Elaborada pelo autor (2022).

Diante disso, a classe de funções de correlação GW rivaliza com a Matérn no sentido de que, segundo Gneiting et al. (2013), as funções do tipo GW permitem a substituição direta da distância euclidiana por grandes distâncias do círculo em uma esfera uni, bi ou tridimensional, isto é, a menor distância entre dois pontos não diretamente opostos na superfície de uma esfera. Além disso, como já dito anteriormente, a classe GW apresenta como característica a geração de matrizes de covariância esparsas, o que é vantajoso quando se trabalha com banco de dados de alta dimensão, haja vista que matrizes esparsas requerem menos espaço de memória, assim, permitem que os algoritmos sejam implementados de forma mais eficiente. Portanto, a motivação de se utilizar uma estrutura de correlação a partir da Wendland está no ganho computacional.

2.3.1 Modelo teórico

Seja o processo $\{\mathbf{Z}(s) : s \in S \subset \mathbb{R}^2\}$. Para sua especificação suponha que o processo satisfaça a condição de estacionariedade de segunda ordem e isotropia, ou seja, $\mu(s) = E(\mathbf{Z}(s)) = \mu$, e que a função de covariância dependa apenas da distância euclidiana entre as localizações. Neste caso, a variância do processo é constante, que aqui identifica-se por $\phi_2 = Var(\mathbf{Z}(s))$.

Portanto, é permitido escrever a covariância como:

$$C(h) = \phi_2 \rho(h) \quad \text{com} \quad h = \|\mathbf{h}\|,$$

em que $\rho(\cdot)$ é a função de correlação. Note que a covariância escrita deste forma tem a mesma representação que a definida em (2.4). Portanto, nesta Seção define-se um modelo paramétrico de covariância substituindo a função de correlação $\rho(\cdot)$ pela correlação do tipo GW. Para tal, considera-se o vetor paramétrico $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)^\top$, em que ϕ_1, ϕ_2 , e ϕ_3 denotam o efeito pepita, contribuição ou variância do processo e parâmetro de escala (neste caso o mesmo que o suporte β).

Então, a partir destas premissas as funções de covariância, correlação e semivariância estão relacionadas, esta última podendo ser escritas como em: (2.8), (2.9), (2.10) e (2.11).

1) Para $\kappa = 0$ define-se a função de covariância como:

$$C(h) = \begin{cases} \phi_1 + \phi_2, & \text{se } h = 0, \\ \phi_2 \left(1 - \frac{h}{\phi_3}\right)^\mu, & \text{se } 0 < h < \phi_3, \\ 0, & \text{se } h \geq \phi_3, \end{cases}$$

função de semivariância

$$\gamma(h) = \begin{cases} 0, & \text{se } h = 0, \\ \phi_1 + \phi_2 \left[1 - \left(1 - \frac{h}{\phi_3}\right)^\mu\right], & \text{se } 0 < h < \phi_3, \\ \phi_1 + \phi_2, & \text{se } h \geq \phi_3, \end{cases} \quad (2.8)$$

e função de correlação

$$\rho(h) = \begin{cases} 1, & \text{se } h = 0, \\ \left(1 - \frac{h}{\phi_3}\right)^\mu, & \text{se } h > 0. \end{cases}$$

2) Para $\kappa = 1$, tem-se a função de covariância dada por

$$C(h) = \begin{cases} \phi_1 + \phi_2, & \text{se } h = 0, \\ \phi_2 \left(1 - \frac{h}{\phi_3}\right)^{\mu+1} \left[\frac{h}{\phi_3}(\mu+1) + 1\right], & \text{se } 0 < h < \phi_3, \\ 0, & \text{se } h \geq \phi_3, \end{cases}$$

função de semivariância

$$\gamma(h) = \begin{cases} 0, & se \quad h = 0, \\ \phi_1 + \phi_2 \left[1 - \left(1 - \frac{h}{\phi_3} \right)^{\mu+1} \left[\frac{h}{\phi_3} (\mu + 1) + 1 \right] \right], & se \quad 0 < h < \phi_3, \\ \phi_1 + \phi_2, & se \quad h \geq \phi_3, \end{cases} \quad (2.9)$$

e função de correlação

$$\rho(h) = \begin{cases} 1, & se \quad h = 0, \\ \left(1 - \frac{h}{\phi_3} \right)^{\mu+1} \left[\frac{h}{\phi_3} (\mu + 1) + 1 \right], & se \quad h > 0. \end{cases}$$

3) Para $\kappa = 2$, tem-se a função de covariância dada por

$$C(h) = \begin{cases} \phi_1 + \phi_2, & se, \quad h = 0, \\ \phi_2 M \left(1 - \frac{h}{\phi_3} \right)^{\mu+2}, & se \quad 0 < h < \phi_3, \\ 0, & se \quad h \geq \phi_3, \end{cases}$$

em que $M = \left(1 + (\mu + 2) \frac{h}{\phi_3} + \frac{h^2}{3\phi_3^2} ((\mu + 2)^2 - 1) \right)$.

Assumindo M definido em $C(h)$ a função de semivariância é

$$\gamma(h) = \begin{cases} 0, & se \quad h = 0, \\ \phi_1 + \phi_2 \left[1 - M \left(1 - \frac{h}{\phi_3} \right)^{\mu+2} \right], & se \quad 0 < h \leq \phi_3, \\ \phi_1 + \phi_2, & se \quad h > \phi_3, \end{cases} \quad (2.10)$$

e função de correlação

$$\rho(h) = \begin{cases} 1, & se \quad h = 0, \\ M \left(1 - \frac{h}{\phi_3} \right)^{\mu+2}, & se \quad h > 0. \end{cases}$$

4) Para $\kappa = 3$, e $q = 1 + \frac{h}{\phi_3}(\mu + 3) + \frac{h^2}{\phi_3^2}(2\mu^2 + 12\mu + 15)\frac{1}{5} + \frac{h^3}{\phi_3^3}(\mu^3 + 9\mu^2 + 23\mu + 15)\frac{1}{15}$

tem-se a função de covariância dada por

$$C(h) = \begin{cases} \phi_1 + \phi_2, & se, \quad h = 0, \\ \phi_2 q \left(1 - \frac{h}{\phi_3} \right)^{\mu+3}, & se \quad 0 < h < \phi_3, \\ 0, & se \quad h \geq \phi_3, \end{cases}$$

a função de semivariância é

$$\gamma(h) = \begin{cases} 0, & \text{se } h = 0, \\ \phi_1 + \phi_2 \left[1 - q \left(1 - \frac{h}{\phi_3} \right)^{\mu+3} \right], & \text{se } 0 < h < \phi_3, \\ \phi_1 + \phi_2, & \text{se } h \geq \phi_3, \end{cases} \quad (2.11)$$

e função de correlação

$$\rho(h) = \begin{cases} 1, & \text{se } h = 0, \\ q \left(1 - \frac{h}{\phi_3} \right)^{\mu+3}, & \text{se } h > 0. \end{cases}$$

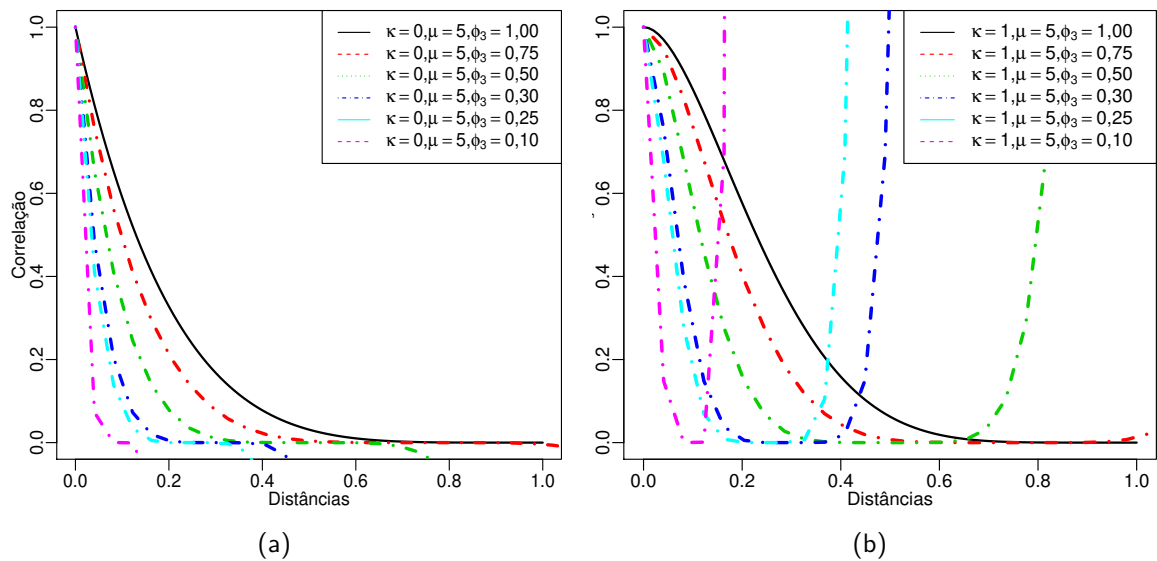
Os gráficos obtidos a partir das funções de covariâncias, correlações, e semivariâncias $(C(h), \rho(h), \gamma(h))$ são nomeados de covariograma, correlograma e semivariograma, respectivamente. Desta forma, e sob a hipótese de estacionariedade de segunda ordem, os três gráficos fornecem a mesma informação descritiva com relação a dependência espacial, embora o semivariograma possua forma invertida com relação aos demais. O semivariograma inicia-se em seu ponto mínimo e cresce para o patamar $(\phi_1 + \phi_2)$, podendo atingi-lo ou não, geralmente assintoticamente. O covariograma e correlograma apresentam o mesmo comportamento, ambos são não-crescentes decaindo para zero com o aumento da distância entre as localizações $(h \rightarrow \infty)$. Distinguindo-se pelo fato do primeiro iniciar em seu ponto máximo (patamar) e o segundo em 1.

Particularmente, o decaimento da covariograma é controlado pelo parâmetro ϕ_3 (suporte). Então, a adaptação das funções de Wendland para análise geoestatística requer a necessidade de dimensionar-lo, podendo ser considerado contínuo, ser estimado, ou mantido constante a depender da natureza do problema. Todavia, tal como o parâmetro de equivalência da família Matérn, em que Zhang (2004) esclarece sobre o problema de identificabilidade, isto é, se a suavidade é fixa e conhecida não é possível que se estime os parâmetros de escala e variância de forma consistente, Bevilacqua et al. (2019) verifica a mesma problemática para este parâmetro.

Diante destas informações, podemos determinar a partir de qual distância a dependência espacial entre as observações torna-se desprezível, isto é, quando avaliada pelo semivariograma obtém-se, geralmente 95% do patamar. Esta distância é comumente chamada de alcance prático ou *range*. Logo, com objetivo de verificar os possíveis valores do suporte que fazem sentido para aplicações da geoestatística, as Figuras 3(a), 3(b), 4(a) e 4(b) ilustram o comportamento do decaimento da função de correlação, aqui sugerida, para vários valores do suporte em seu

domínio de definição. Sem perda de generalidade os parâmetros de positividade e variância μ e ϕ_2 são considerados constantes e iguais a 5 e 1, respectivamente. Percebe-se, em todos os casos, o decaimento da curva para valores do suporte próximos de zero ocorre de forma mais acentuada, sendo notável a ocorrência de possíveis correlações negativas. Este comportamento não é aceitável para os pressupostos aqui adotados, caso contrário, teríamos a existência de semivariâncias negativas. O comportamento parabólico, percebido nos gráficos 3(b) e 4(b), possivelmente está relacionado a pequena variação em curtas distâncias. Nota-se ainda, que quando o suporte se aproxima da origem a área de correlação espacial tende a diminuir, ocorrendo um processo estritamente inverso quando o suporte se aproxima do extremo do domínio (1). Portanto, fica claro que para questões práticas, sob as suposições de validade das funções de covariância, correlação e semivariância, é indicado que se tome o suporte nas vizinhanças de 1, podendo ser maior ou menor, contudo, deve-se observar a ordem de grandeza dos dados.

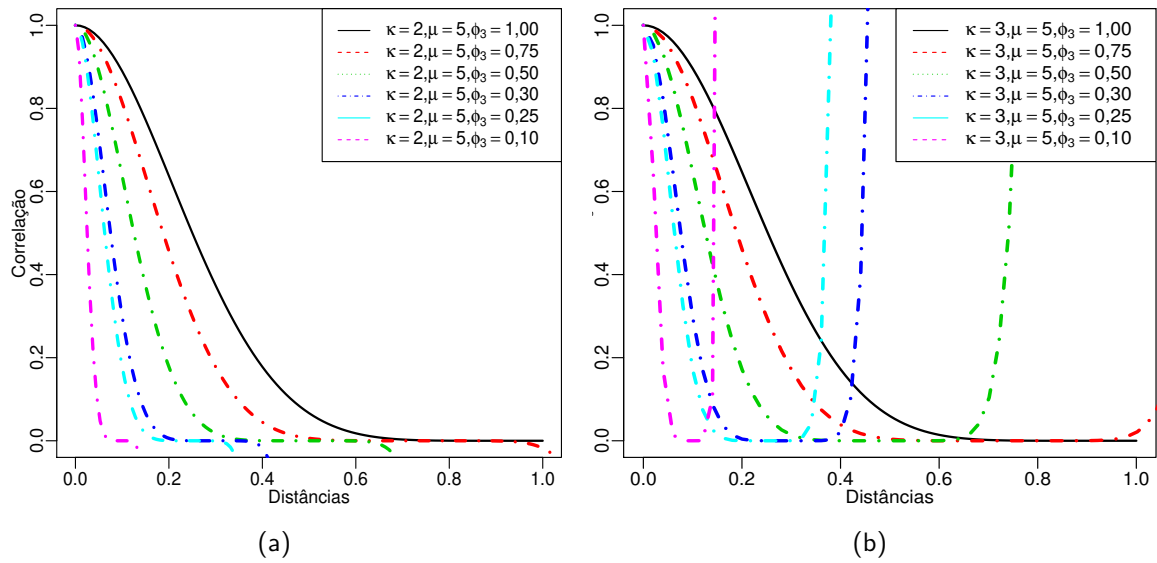
Figura 2 – Correlação de Wendland Generalizada para suavização $\kappa = 0$, em (a) e para $\kappa = 1$, em (b), para vários valores de ϕ_3 , com parâmetro que define a positividade da função de covariância $\mu = 5$ em ambos os casos.



Fonte: Elaborada pelo autor (2022).

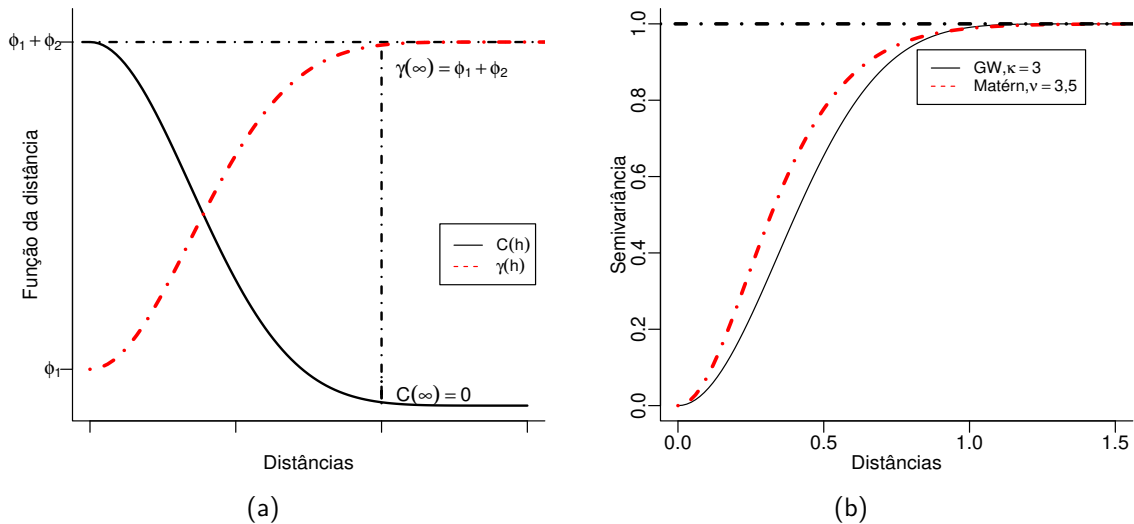
Assim sendo, e tendo em consideração a equivalência entre as famílias de covariância de GW e Matérn, tal como exposto na Tabela 1, ainda, dada a relação entre a função de covariância e semivariância, sob suposição de estacionariedade de segunda ordem, na Figura 4 é apresentado os gráficos da relação entre as funções $\gamma(\mathbf{h}) \times h$ e $C(\mathbf{h}) \times h$ para a família GW, em 5(a), e equivalência das funções de semivariâncias das famílias GW e Matérn, 5(b) para suavização $\kappa = 3$ e $\nu = 3,5$, respectivamente.

Figura 3 – Correlação de Wendland Generalizada para suavização $\kappa = 2$, em (a) e para $\kappa = 3$, em (b), para vários valores de ϕ_3 , com parâmetro que define a positividade da função de covariância $\mu = 5$ em ambos os casos.



Fonte: Elaborada pelo autor (2022).

Figura 4 – Gráfico da relação entre a função de correlação e semivariância da Wendland Generalizada para suavização $\kappa = 3, 5$ (a). Funções de Semivariância para família de Wendland Generalizada e Matérn com suavização $\nu = 3, 5$. Com alcances prático aproximadamente iguais a 0,82 e 0,76 ($\phi_3 = 1,6$ e 0,11) para Wendland e Matérn, respectivamente. Em que $g(\cdot)$ é função que varia para cada caso.



Fonte: Elaborada pelo autor (2022).

Atente que, tal qual o comportamento da família de covariância Matérn, em particular para parâmetros de escala equivalentes, a família de covariância GW evidencia um comportamento monotônico não decrescente e não crescente, respectivamente, para as funções de semivariância e correlação, ver Figura 5(a). Ainda, seu comportamento explicita sua competitividade

com a família Matérn, ver Figura 5(b). Portanto, em particular, fica claro que a função de semivariância da família GW atinge um patamar assintótico, portanto, o alcance prático pode ser obtido, aproximadamente, como função do suporte ϕ_3 para cada caso do parâmetro de suavização κ , aderido neste trabalho.

2.3.2 Alcance prático e Área de Continuidade Espacial para as família de covariância Wendland generalizada e Matérn

A velocidade com que $\rho(h) \rightarrow 0$ quando a distância aumenta reflete a extensão física da correlação espacial no processo. Assim, se $\rho(h) = 0$ para uma distância h maior que um número real finito, $a > 0$, então, o intervalo $(0, a]$ é nomeado de intervalo de correlação espacial. No entanto, em geral, a função de correlação $\rho(h)$, assume valor nulo apenas assintoticamente, e neste caso, o intervalo de correlação espacial é indefinido. À vista disto, é comum na literatura adotar a convenção $\rho(a) = 0,05$, melhor dizendo, $\gamma(a) = \phi_1 + 0,95\phi_2$, para se determinar o alcance prático.

Desta forma, admitindo algumas suposições para o parâmetro μ - o qual determina a positividade da matriz de correlação. Assim, aqui, determina-se as expressões para a , em função de ϕ_3 , para os modelos de GW com parâmetros de suavização $\kappa = 0, 1, 2$ e 3 , ver Equações (2.12), (2.13), (2.14) e (2.15). Em particular, como visto na Seção 2.3, a correlação para família GW assume valor zero para distâncias maiores que seu suporte. Desta maneira, $a = g(\phi_3)$. Assim, o alcance fica definido como segue.

i) Para suavização $\kappa = 0$, tem -se que $\mu = 2$. Então,

$$a = (1 - 0,05^{1/\mu}) \phi_3 = 0,776\phi_3, \quad (2.12)$$

ii) Para suavização $\kappa = 1$, tem -se que $\mu = 3$. Então,

$$a = 0,657\phi_3, \quad (2.13)$$

iii) Para suavização $\kappa = 2$, tem -se que $\mu = 4$. Então,

$$a = 0,573\phi_3, \quad (2.14)$$

iv) Para suavização $\kappa = 3$, tem -se que $\mu = 5$. Então,

$$a = 0,514\phi_3, \quad (2.15)$$

em todos os casos adota-se $\mu = \frac{d+1}{2} + \kappa + 0,5$, $d = 2$.

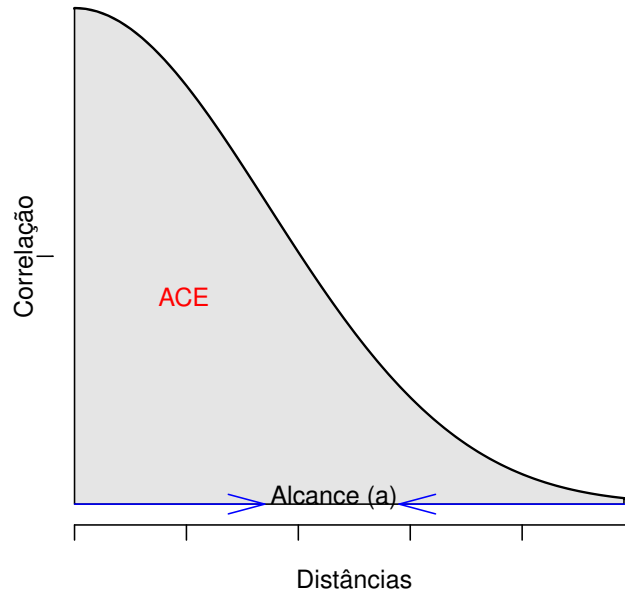
Note que, quando se utiliza esta abordagem para famílias de covariâncias que apresentam parâmetros de suavização, aumenta-se o nível de complexidade para encontrar uma expressão válida para o alcance prático em função do parâmetro de escala ϕ_3 (particularmente o suporte), uma vez que é preciso resolver uma equação não linear. Em particular, para o caso da GW, a resolução das equações não lineares produzem raízes reais e complexas. Sendo simples para $\kappa = 0$. Assim, como o alcance é definido para um número real positivo, considera-se as raízes reais positivas.

2.3.3 Área de Correlação Espacial e Fator do Modelo

Ao considerar o semivariograma e o ajuste para os modelos teóricos, há possibilidade de uma avaliação visual da dependência espacial. Desta forma, faz-se necessário a obtenção de algum valor numérico que mensure esta dependência. Em geral, na literatura especializada utiliza-se uma medida que relaciona o efeito pepita e o patamar. Nesta conjuntura, foi proposta por Seidel e Oliveira (2014) uma nova medida, nomeada Índice de Dependência Espacial (IDE), que relacione não somente o efeito pepita e o patamar mas também toda a geometria do semivariograma. Apesar dos autores estabelecerem uma expressão para o IDE (ver Equação (2.16)), os termos envolvidos foram determinados apenas para os modelos Esférico, Exponencial e Gaussiano. Diante disso, nesta Seção são calculados os elementos envolvidos neste novo índice para família de covariância Wendland e Matérn.

Ao modelar determinado fenômeno sob as hipóteses de estacionariedade de segunda ordem, o semivariograma e covariograma se relacionam por $C(h) = 1 - \frac{\gamma(h)}{C(0)}$. Desta forma, os elementos para o cálculo do IDE podem ser obtidos pelo cálculo da área da zona de influência ou correlação espacial (Área de continuidade (ou correlação) espacial (ACE)), ver Figura 5.

Figura 5 – Imagem ilustrativa de uma região de correlação espacial para um correlograma pertencente a família de Wendland, definida a partir da Equação (2.2) com suavização $\kappa = 3$.



Fonte: Elaborada pelo autor (2022)

A partir do cálculo da área da zona de influência, os autores Seidel e Oliveira (2014) estabeleceram o novo índice de dependência como sendo dado pela Equação (2.16).

$$IDE(\%) = FM \times \left(\frac{\hat{\phi}_2}{\hat{\phi}_1 + \hat{\phi}_2} \right) \times \left(\frac{a}{qMD} \right) \times 100, \quad (2.16)$$

em que $Fator_{domodelo}(FM)$ indica a fator que mede a força da dependência espacial do modelo, quanto maior seu valor, maior será a dependência espacial do modelo. O valor de FM é obtido pelo cálculo da área de correlação espacial do modelo. Os termos ϕ_1 , ϕ_2 e a são o efeito pepita, a contribuição e o alcance prático, respectivamente, enquanto o termo qMD denota a fração (q) resultante da máxima distância ($Máximadistância(MD)$) entre os pontos amostrados. Os autores esclarecem que a fração $\frac{a}{qMD}$ deve estar entre zero e 1 (um), na hipótese de se obter um valor maior que 1 seu valor é fixado em 1. No trabalho desenvolvido pelos autores é apresentado o fator do modelo (FM) para os casos dos modelos Esférico, Exponencial e Gaussiano iguais a $\frac{3}{8}$, $\frac{e^3+1}{3e^3}$ e $\frac{\sqrt{\pi}erf(\sqrt{3})}{2\sqrt{3}}$, respectivamente, que podem ser expressos em uma escala decimal pelos valores 0,375, 0,317, e 0,504.

Portanto, aqui se determina as expressões para os $ACE's$ da família GW, denotando por ACE_{κ} – Área de Correlação Espacial para família GW – com $\kappa = 0, 1, 2$ e 3 os parâmetros de suavização utilizados. Como resultado se obtém o Fator do Modelo para família GW, FM_{κ} . O

desfecho, operando com a relação entre correlação e semivariância, é ilustrado nas Equações (2.17), (2.19), (2.20) e (2.21).

$$\begin{aligned}
 ACE_0 &= \int_0^a \left[1 - \left(\frac{\phi_1}{\phi_1 + \phi_2} + \frac{\phi_2}{\phi_1 + \phi_2} \left(1 - \left(1 - \frac{h}{\phi_3} \right)^\mu \right) \right) \right] dh \\
 &= \left[h - \frac{\phi_1}{\phi_1 + \phi_2} h - \frac{\phi_2}{\phi_1 + \phi_2} h \right]_0^a + \frac{\phi_2}{\phi_1 + \phi_2} \frac{a}{0,776} \int_{0,224}^1 v^\mu dv \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} a \times 0,425,
 \end{aligned} \tag{2.17}$$

em que no processo de integração de (2.17) foi aplicada a mudança de variável $v = 1 - \frac{h}{\phi_3}$, $\mu = 2$, definido em (2.3), e a relação (2.12). Assim, tem-se o fator do modelo para suavização $\kappa = 0$ na Equação (2.18).

$$FM_0 = 0,425. \tag{2.18}$$

Como os termos $1 - \left(\frac{\phi_1}{\phi_1 + \phi_2} + \frac{\phi_2}{\phi_1 + \phi_2} \right)$ do integrando em (2.17) se cancelam, então, por simplicidade, nos demais resultados os mesmos serão omitidos. Sendo assim, considerando o parâmetro μ iguais a 3, 4 e 5, para os casos de $\kappa = 1, 2, 3$, obtém-se os resultados para o ACE_1, ACE_2 e ACE_3 em (2.19), (2.20) e (2.21), respectivamente.

$$\begin{aligned}
 ACE_1 &= \frac{\phi_2}{\phi_1 + \phi_2} \int_0^a \left(1 - \frac{h}{\phi_3} \right)^4 \left(4 \frac{h}{\phi_3} + 1 \right) dh \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} \frac{a}{0,657} \int_{0,343}^1 v^4 (4(1-v) + 1) dv \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} a \times 0,502,
 \end{aligned} \tag{2.19}$$

$$\begin{aligned}
 ACE_2 &= \frac{\phi_2}{\phi_1 + \phi_2} \int_0^a \left(1 - \frac{h}{\phi_3} \right)^6 \left(1 + 6 \frac{h}{\phi_3} + \frac{35}{3} \frac{h^2}{\phi_3^2} \right) dh \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} \frac{a}{0,573} \int_{0,427}^1 \left(7 - 6v + \frac{35}{3} (1 - 2v + v^2) \right) v^6 dv \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} a \times 0,511,
 \end{aligned} \tag{2.20}$$

$$\begin{aligned}
 ACE_3 &= \frac{\phi_2}{\phi_1 + \phi_2} \int_0^a \left(1 - \frac{h}{\phi_3} \right)^8 \left(1 + 8 \frac{h}{\phi_3} + 25 \frac{h^2}{\phi_3^2} + 32 \frac{h^3}{\phi_3^3} \right) dh \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} \frac{a}{0,514} \int_{0,486}^1 (66v^8 - 154v^9 + 121v^{10} - 32v^{11}) dv \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} a \times 0,512,
 \end{aligned} \tag{2.21}$$

em que no processo de integração das Equações (2.19), (2.20) e (2.21) é utilizada as mudanças de variáveis $v = 1 - \frac{h}{\phi_3}$ e $\frac{h}{\phi_3} = 1 - v$.

Analogamente, utilizando as relações de equivalência constantes na Tabela 1, obtém-se as - ACE_{ν_0} , ACE_{ν_1} , ACE_{ν_2} e ACE_{ν_3} - bem como os - FM_{ν_0} , FM_{ν_1} , FM_{ν_2} e FM_{ν_3} - para indicar a Área de Correlação Espacial e o Fator do Modelo da família Matérn, com as correspondentes equivalências de acordo com o parâmetro de suavização, $\nu = 0,5$; $\nu = 1,5$; $\nu = 2,5$; e $\nu = 3,5$, nessa ordem. Para os valores do alcance da família Matérn, De BASTIANI (2012) descreve $a = 3\phi_3$, $a = 4,75\phi_3$ e $a = 5,92\phi_3$, para os parâmetros de suavização $\nu = 0,5$; $\nu = 1,5$; e $\nu = 2,5$, respectivamente. Para o caso de $\nu = 3,5$ foi aqui calculado e obtido $a = 6,76\phi_3$.

Assim sendo, as expressões para ACE e os respectivos valores para o FM da família Matérn estão determinados pelas Equações (2.22), (2.24), (2.26), e (2.28).

$$\begin{aligned} ACE_{\nu_0} &= \int_0^{a=3\phi_3} \left[1 - \left(\frac{\phi_1}{\phi_1 + \phi_2} + \frac{\phi_2}{\phi_1 + \phi_2} (1 - e^{-h/\phi_3}) \right) \right] dh \\ &= \frac{\phi_2}{\phi_1 + \phi_2} \left[-\phi_3 e^{-h/\phi_3} \right]_0^{a=3\phi_3} \\ &= \frac{\phi_2}{\phi_1 + \phi_2} a \left(\frac{e^3 - 1}{3e^3} \right), \end{aligned} \quad (2.22)$$

então, o fator do modelo da família Matérn para suavização $\nu = 0,5$ é dado pela Equação (2.23).

$$FM_{\nu_0} = \frac{e^3 - 1}{3e^3} \approx 0,317, \quad (2.23)$$

em que, como esperado, é o mesmo valor para o modelo Exponencial, haja vista que, em particular, coincide com o modelo Matérn em $\nu = 0,5$.

A área de continuidade espacial para o caso de $\nu = 1,5$, tem-se a Equação (2.24).

$$\begin{aligned} ACE_{\nu_1} &= \frac{\phi_2}{\phi_1 + \phi_2} \int_0^{a=4,75\phi_3} \left(1 + \frac{h}{\phi_3} \right) e^{-h/\phi_3} dh \\ &= \frac{\phi_2}{\phi_1 + \phi_2} \int_0^{a=4,75\phi_3} \left(e^{-h/\phi_3} + \frac{h}{\phi_3} e^{-h/\phi_3} \right) dh \\ &= \frac{\phi_2}{\phi_1 + \phi_2} \left[-e^{-h/\phi_3} (2\phi_3 + h) \right]_0^{a=4,75\phi_3} \\ &= \frac{\phi_2}{\phi_1 + \phi_2} a \left(\frac{2e^{4,75} - 6,75}{4,75e^{4,75}} \right). \end{aligned} \quad (2.24)$$

Note que na resolução de (2.24) aplicou-se a integração por partes, $u = h$ e $dv = e^{-h/\phi_3}$. Posto isto, o fator do modelo, pertencente a família Matérn para o caso da suavização $\nu = 1,5$, é dado pela Equação (2.25).

$$FM_{\nu_1} = \frac{2e^{4,75} - 6,75}{4,75e^{4,75}} \approx 0,409. \quad (2.25)$$

Na hipótese de $\nu = 2,5$, tem-se a área de continuidade espacial obtida a partir da Equação (2.26).

$$\begin{aligned}
 ACE_{\nu_2} &= \frac{\phi_2}{\phi_1 + \phi_2} \int_0^{a=5,92\phi_3} \left(1 + \frac{h}{\phi_3} + \frac{h^2}{3\phi_3^2} \right) e^{-h/\phi_3} dh \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} \left[-\phi_3 e^{-h/\phi_3} - 2\phi_3 e^{-h/\phi_3} - \frac{e^{-h/\phi_3} (h^2 + 2\phi_3(h + a))}{3\phi_3} \right]_0^{a=5,92\phi_3} \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} a (0,446 - 6,712e^{-5,92}), \tag{2.26}
 \end{aligned}$$

em que utilizou-se os resultados obtidos nas Equações, (2.22), (2.24), e uma integração por partes do tipo $u = h^2$ e $dv = e^{-h/\phi_3}$. Então, neste caso, o fator do modelo é dado na Equação (2.27).

$$FM_{\nu_2} = 0,45 - 6,712e^{-5,92} \approx 0,439. \tag{2.27}$$

Para o caso de $\nu = 3,5$ tem-se a Equação (2.28).

$$\begin{aligned}
 ACE_{\nu_3} &= \frac{\phi_2}{\phi_1 + \phi_2} \int_0^{a=6,76\phi_3} \left(1 + \frac{h}{2\phi_3} + \frac{2h^2}{5\phi_3^2} + \frac{h^3}{15\phi_3^3} \right) e^{-h/\phi_3} dh \\
 &= \frac{\phi_2}{\phi_1 + \phi_2} \left[-\phi_3 (e^{-a/\phi_3} - 1) + \frac{\phi_3 - e^{-a/\phi_3}(a + \phi_3)}{2} \right. \\
 &\quad + \frac{2}{5\phi_3} (2\phi_3^2 - e^{-a/\phi_3}(a^2 + 2\phi_3(\phi_3 + a))) \\
 &\quad \left. + \frac{-a^3 e^{-a/\phi_3} - 3\phi_3 (e^{-a/\phi_3} (\frac{a^2}{\phi_3^2} + \frac{2a}{\phi_3} + 2) - 2)}{15\phi_3^2} \right] \\
 &\approx \frac{\phi_2}{\phi_1 + \phi_2} a \times 0,444, \tag{2.28}
 \end{aligned}$$

em que aplicou-se, além dos resultados obtidos por (2.22), (2.24) e (2.26), uma mudança de variável $y = -h/\phi_3$ na parte correspondente ao integrando $\frac{h^3}{15\phi_3^3} e^{-h/\phi_3}$, em seguida aplica-se a integração por partes com $u = y^3$ e $dv = e^y$.

Logo, os FM's obtidos para famílias GW e Matérn considerando vários valores de suavização estão anotados na Tabela 2. Assim sendo, é possível calcular o nível de dependência espacial a partir do modelo utilizado, salientando que este nível varia entre zero e o FM, este, por sua vez, caracteriza a força da dependência espacial. Assim, quanto maior for seu valor, maior será a força de caracterização da dependência espacial. Assim sendo, dos resultados obtidos é possível concluir que os modelos com estrutura de covariância a partir da Wendland são melhores para caracterizar a dependência espacial, perdendo apenas nos casos de suavização κ iguais a 0 e 1 para o modelo Gaussiano com valor de FM igual a 0,504 já obtido na literatura.

Tabela 2 – Fator do Modelo para Wendland e Matérn.

$(\kappa; \nu)$	FM - Wendland	FM - Matérn
(0; 0, 5)	0,425	0,317
(1; 1, 5)	0,502	0,409
(2; 2, 5)	0,511	0,439
(3; 3, 5)	0,512	0,444

Fonte: Elaborada pelo autor (2022).

Intuitivamente, em analogia com o índice de dependência espacial definido em Biondi, Myers e Avery (1994), e como bem destaca Seidel e Oliveira (2015), é permissível construir uma escala de classificação para o *IDE* a partir da medida

$$FM \times \left(\frac{\hat{\phi}_2}{\hat{\phi}_1 + \hat{\phi}_2} \right),$$

neste caso o *FM* é visto como um fator de proporcionalidade. Com isso, estendemos tais resultados para as famílias de covariâncias Wendland e Matérn, com resultados obtidos nas Tabelas classificações 3 e 4.

Tabela 3 – Escala de classificação segundo a dependência espacial entre as amostras para família Wendland.

$\kappa = 0$	$IDE \leq 11\%$	fraca dependência espacial
	$11\% < IDE \leq 32\%$	moderada dependência espacial
	$IDE > 32\%$	forte dependência espacial
$\kappa = 1$	$IDE \leq 13\%$	fraca dependência espacial
	$13\% < IDE \leq 38\%$	moderada dependência espacial
	$IDE > 38\%$	forte dependência espacial
$\kappa = 2$	$IDE \leq 13\%$	fraca dependência espacial
	$13\% < IDE \leq 38\%$	moderada dependência espacial
	$IDE > 38\%$	forte dependência espacial
$\kappa = 3$	$IDE \leq 13\%$	fraca dependência espacial
	$13\% < IDE \leq 38\%$	moderada dependência espacial
	$IDE > 38\%$	forte dependência espacial

Fonte: Elaborada pelo autor (2022).

Tabela 4 – Escala de classificação segundo a dependência espacial entre as amostras para família Matérn.

$\nu = 0,5$	$IDE \leq 8\%$	fraca dependência espacial
	$8\% < IDE \leq 24\%$	moderada dependência espacial
	$IDE > 24\%$	forte dependência espacial
$\nu = 1,5$	$IDE \leq 10\%$	fraca dependência espacial
	$10\% < IDE \leq 31\%$	moderada dependência espacial
	$IDE > 31\%$	forte dependência espacial
$\nu = 2,5$	$IDE \leq 11\%$	fraca dependência espacial
	$11\% < IDE \leq 33\%$	moderada dependência espacial
	$IDE > 33\%$	forte dependência espacial
$\nu = 3,5$	$IDE \leq 11\%$	fraca dependência espacial
	$11\% < IDE \leq 33\%$	moderada dependência espacial
	$IDE > 33\%$	forte dependência espacial

Fonte: Elaborada pelo autor (2022).

Posto isto, na seção 2.4 dá-se início ao processo de estimação dos parâmetros para família de covariância GW. No caso da família Matérn, os parâmetros são estimados por funções já disponíveis no *Software R*.

2.4 ESTIMAÇÃO DOS PARÂMETROS

O processo de estimação paramétrica para estrutura de covariância espacial se dá por método de Máxima Verossimilhança (Máxima verossimilhança (ML)). No entanto, como é sabido, este método exige além do conhecimento da função de verossimilhança a informação de um vetor de parâmetros para iniciar o processo. Diante disso, por simplicidade, sugere-se determinar o vetor de parâmetros iniciais pelo método de estimação por Mínimos Quadrados Ordinários (Mínimos quadrados ordinários (OLS)).

2.4.1 Método dos Mínimos Quadrados Ordinários

Considere o processo estocástico intrinsecamente estacionário $Z = \{Z(\mathbf{s}_i) : \mathbf{s}_i \in \mathbf{S} \subset \mathbb{R}^d, d \geq 1\}$. Seja um vetor aleatório de dimensão $n \times 1$, $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))^T$, com $Z(\mathbf{s}_i)$ definindo as n 's observações do processo estocástico nas localizações distintas $\mathbf{s}_i, i = 1, \dots, n$. Posto isto, como é bem definido na literatura (ver por exemplo Cressie (1993), Uribe-Opazo, Borssoi e Galea (2012)), as n 's observações do processo são geradas por um modelo linear

especial definido como:

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$

em que $\mu(\mathbf{s}_i)$ denota a componente dos efeitos fixos do processo, isto é, o termo determinístico. O termo $\epsilon(\mathbf{s}_i)$ indica o erro do processo, ou seja, a parte estocástica, em que assume-se ter média zero, $E(\epsilon(\mathbf{s}_i)) = 0$, e variância finita determinada por uma função de covariância $Cov(\epsilon(\mathbf{s}_i), \epsilon(\mathbf{s}_j)) = C(\mathbf{s}_i, \mathbf{s}_j)$, $i, j = 1, \dots, n$, sendo a função de covariância $C(\mathbf{s}_i, \mathbf{s}_j)$ especificada por um vetor de parâmetros $\phi = (\phi_1, \dots, \phi_q)^\top$. Ainda, dado um conjunto $\{x_j(\mathbf{s}_i), j = 1, \dots, p, i = 1, \dots, n\}$, de variáveis explicativas, a média do processo estocástico é definida por:

$$\mu(\mathbf{s}_i) = \sum_{j=1}^p x_j(\mathbf{s}_i) \beta_j, \quad (2.29)$$

em que os β_j 's indicam o j -ésimo parâmetro desconhecido a ser estimado. Podendo ser reescrito em forma matricial como:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.30)$$

em que $\mathbf{X}\boldsymbol{\beta}$ é a média do processo definida em (2.29).

Estimar o vetor de parâmetros $\boldsymbol{\beta}$ em (2.30), exige que se minimize a soma dos quadrados dos erros, ou seja, determinar os valores de $\boldsymbol{\beta}$ que minimizam a equação (2.31).

$$\begin{aligned} \sum_{i=1}^n \hat{\epsilon}^2(\mathbf{s}_i) &= \sum_{i=1}^n \left(Z(\mathbf{s}_i) - \sum_{j=1}^p x_j(\mathbf{s}_i) \beta_j \right)^2, \\ \Rightarrow \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}, \end{aligned} \quad (2.31)$$

que é um estimador usual e não viesado para $\boldsymbol{\beta}$.

Nesta conjuntura, fundamenta-se nas ideias desenvolvidas em Cressie (1985). Com isso, a estimação dos parâmetros de um modelo paramétrico de covariância é feita a partir do ajuste do modelo teórico de um semivariograma/variograma experimental. Este, em geral é representado por:

$$2\gamma(\mathbf{h}) = Var [Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})], \forall \mathbf{s}, \mathbf{s} + \mathbf{h} \in \mathbb{R}^d, \quad (2.32)$$

ou

$$\gamma(\mathbf{h}) = \frac{1}{2} Var (Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})), \forall \mathbf{s}, \mathbf{s} + \mathbf{h} \in \mathbb{R}^d, \quad (2.33)$$

em que (2.32) e (2.33) representam a função de variância e semivariância, respectivamente. Ambas produzem resultados similares, e são utilizados para mensurar a dependência espacial entre pontos separados por uma distância h .

Aqui, é considerado que o campo aleatório $Z(\cdot)$ é intrinsecamente estacionário, $\mathbb{R}^d = \mathbb{R}^2$, e isotrópico em n locais amostrados $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$, ver (CRESSIE, 1993). Dessa maneira, supõe-se ainda, que o variograma assuma uma forma paramétrica conhecida, de tal modo que valham as seguintes propriedades:

i) $E[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] = 0$

ii) $E[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2] = 2\gamma(h_{ij}; \phi)$,

no qual $h_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$ indica a distância euclidiana entre os locais $\mathbf{s}_i$ e $\mathbf{s}_j$. Ante essas hipóteses, pode-se trabalhar no processo de estimação para o semivariograma. Dentre os estimadores disponíveis na literatura, um dos mais utilizados por sua simplicidade e de fácil implementação, obtido a partir do método dos momentos, proposto por Matheron em 1963 comumente denominado estimador clássico de Matheron, (CRESSIE, 1985), o qual é apresentado na equação (2.34),

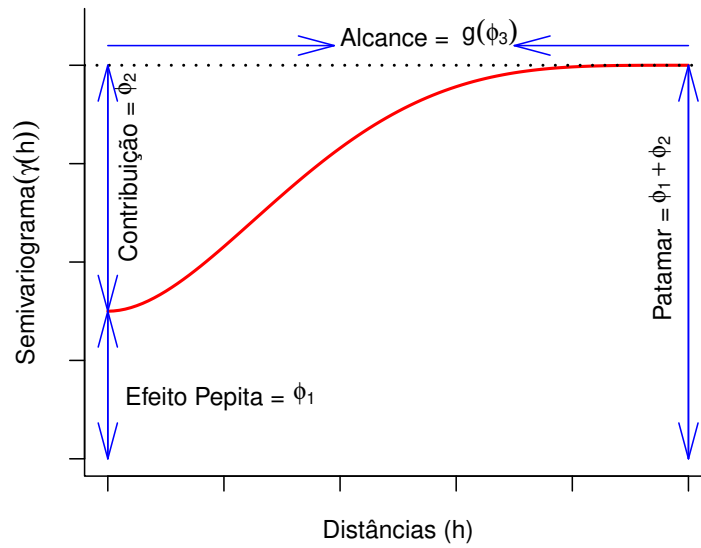
$$\gamma(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{|N(\mathbf{h})|} (Z(\mathbf{s}_i + \mathbf{h}) - Z(\mathbf{s}_i))^2, \quad (2.34)$$

em que $|N(\mathbf{h})|$ denota o número de elementos pertencentes ao conjunto dos pares de pontos com distância h , $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| = h\}$.

Desta feita, e tal como define-se em (2.31), a estimação dos parâmetros em modelos geoestatísticos, consiste em se obter um vetor de parâmetros $\hat{\theta} \in \mathbb{R}^{p+q}$, estimador de $\theta \in \mathbb{R}^{p+q}$ com $\theta = (\beta^\top, \phi^\top)^\top$ e p finito, tal que a partir de pares de vetores $(\mathbf{h}_i, \mathbf{Z}_i)$, amostrados como consequência do mapeamento $\gamma(\mathbf{h})$, com $\mathbf{Z}_i = \gamma(\mathbf{h}_i) + \epsilon_i$ e dado o modelo paramétrico $\gamma(\mathbf{h}; \hat{\phi})$ de tal modo que $dist(\gamma(\mathbf{h}), \gamma(\mathbf{h}; \hat{\phi})) \leq dist(\gamma(\mathbf{h}), \gamma(\mathbf{h}; \phi))$, em que o operador $dist(\cdot)$ mede a distância entre duas funções em um subespaço do $\mathbb{R}^d$, enquanto o vetor ϵ_i representa o erro aleatório, usualmente assumido ser de média zero e variância constante. Destarte, a solução desse problema, caso exista, é a melhor aproximação e depende diretamente da classe de funções a qual $\gamma(\mathbf{h}; \phi)$ pertence.

Consequentemente, para um número finito n de dados amostrados em posições distintas $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$, e definido um modelo paramétrico $\gamma(\mathbf{h}; \phi)$, a distância entre a função a ser aproximada e sua aproximação é sempre uma função dependente do vetor de parâmetros ϕ . Na figura 6 é ilustrado o gráfico de semivariograma $\gamma(\mathbf{h})$ *versus* $\mathbf{h}$, em que se define o vetor de parâmetros $\phi = (\phi_1, \phi_2, \phi_3)^\top$, com $q = 3$, para compor a estrutura de dependência espacial a qual se aborda neste trabalho.

Figura 6 – Gráfico ilustrativo do semivariograma, $\gamma(\mathbf{h})$ versus $\mathbf{h}$, para família Wendland com parâmetro de suavização $\kappa = 1$.



Fonte: Elaborada pelo autor (2022).

A partir do padrão visualizado na Figura 6 é possível perceber, mesmo que intuitivamente, o que se espera da diferença $Z(s_i + \mathbf{h}) - Z(s_i)$ utilizada na expressão (2.34), isto é, a medida que a distância h decresce, seus valores decrescem na mesma proporção. Note, ainda, que os valores para variância no gráfico 6 são obtidos por meio da função de variância avaliada na distância entre duas localizações, e não nos dados amostrados, o que faz sentido, haja vista que é assumida a hipótese de estacionariedade intrínseca.

Posto isso, prossegue-se no processo de estimação para os modelos GW por OLS considerando que, em Cressie (1985), é adotada como medida de distância a norma euclidiana e define o funcional

$$Q(\mathbf{h}; \boldsymbol{\theta}) = \sum_{i,j=1}^m (\gamma(h_{ij}) - \gamma(h_{ij}; \boldsymbol{\phi}))^2, \quad (2.35)$$

em que m denota a quantidade de conjuntos com pares de pontos associados a uma certa distância h , pré-definida, do semivariograma experimental; $\gamma(h_{ij})$, o valor estimado da função de semivariância; e $\gamma(h_{ij}; \boldsymbol{\phi})$, o valor estimado pelo modelo ajustado.

Nesse cenário, e como é sabido que o método dos mínimos quadrados ordinários consiste em uma técnica de se determinar os parâmetros ótimos, então, com esta prerrogativa, este método esclarece que é preciso minimizar o funcional (2.35). Para tanto, existe na literatura uma diversidade de técnicas de minimização de funções, dentre elas derivar-la com relação a cada um de seus parâmetros, igualar-las a zero, logo após resolver o sistema de equações obtido. Todavia, em sua maioria o funcional (2.35) caracteriza uma função não-linear, em

consequência, o sistema proveniente será um sistema de equações não - lineares que em geral não admite forma fechada.

Desse modo, para resolver problemas com tal característica é preciso recorrer aos métodos de aproximação numérica. Dentre estes, os mais populares na literatura estão os métodos do tipo Quasi - Newton, com destaque para os algoritmo de otimização não linear BFGS (Broyden-Fletcher-Goldfarb-Shanno) e L-BFGS que é mesmo algoritmo BFGS, porém, com uma quantidade limitada de memória, (PRESS et al., 2007; SANTOS, 2019).

Diante disso, e em conformidade com os critérios de positividade dados por (2.3), minimiza-se (2.35) utilizando cada um dos modelos de semivariância enunciados na subseção 2.3.1. Então, definindo os vetores $\gamma = (\gamma_1(\mathbf{h}), \dots, \gamma_m(\mathbf{h}))$ e $\hat{\gamma} = (\gamma_1(\mathbf{h}; \phi), \dots, \gamma_m(\mathbf{h}; \phi))$ para representar o vetor dos valores do semivariograma experimental e o vetor dos valores obtidos pelo modelo paramétrico, respectivamente, e substituindo os modelos de semivariância representados na subseção 2.3.1, obtém-se $Q(\mathbf{h}, \phi)_\kappa$, em que $\kappa = 0, 1, 2, 3$ indicam a ordem do parâmetro de suavização para o modelo de semivariância associado a GW:

$$\begin{aligned} Q(\mathbf{h}, \phi)_0 &= \sum_{i,j=1}^m (\gamma(h_{ij}) - \hat{\gamma}(h_{ij}))^2 \\ &= \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left[1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^2 \right] \right]^2, \end{aligned}$$

$$\begin{aligned} Q(\mathbf{h}, \phi)_1 &= \sum_{i,j=1}^m (\gamma(h_{ij}) - \hat{\gamma}(h_{ij}))^2 \\ &= \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left[1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^4 \left(4 \frac{h_{ij}}{\phi_3} + 1 \right) \right] \right]^2, \end{aligned}$$

$$\begin{aligned} Q(\mathbf{h}, \phi)_2 &= \sum_{i,j=1}^m (\gamma(h_{ij}) - \hat{\gamma}(h_{ij}))^2 \\ &= \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left[1 - \left(1 + 6 \frac{h_{ij}}{\phi_3} + 35 \frac{h_{ij}^2}{3\phi_3^2} \right) \left(1 - \frac{h_{ij}}{\phi_3} \right)^6 \right] \right]^2, \end{aligned}$$

$$\begin{aligned} Q(\mathbf{h}, \phi)_3 &= \sum_{i,j=1}^m (\gamma(h_{ij}) - \hat{\gamma}(h_{ij}))^2 \\ &= \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left[1 - \left(1 + \frac{8h_{ij}}{\phi_3} + \frac{25h_{ij}^2}{\phi_3^2} + \frac{32h_{ij}^3}{\phi_3^3} \right) \left(1 - \frac{h_{ij}}{\phi_3} \right)^8 \right] \right]^2, \end{aligned}$$

que são as somas dos quadrados das diferenças para os casos em que κ são iguais a 0, 1, 2 e 3, respectivamente, e ϕ_3 tem o papel do parâmetro de escala, β , em (2.4). Isso posto, procede-se

com a minimização uma a uma. Para tal, é preciso resolver os sistemas de equações gerados a partir dos cálculo de suas respectivas derivadas parciais, ou seja, $\frac{\partial Q(\mathbf{h}, \phi)_0}{\partial \phi_i}$, $i = 1, 2, 3$, estão expressas no Apêndice B.

Por conseguinte, os sistemas de equações gerados, em geral, podem ser resolvidos por algum dos métodos do tipo Quasi - Newton, ver por exemplo Santos (2019), Cordeiro e Demétrio (2013). Neste trabalho é utilizado o método de estimação L-BFGS com a função *optim()* disponível no *Software* (R Core Team, 2020).

2.4.2 Método de Máxima Verossimilhança

Para que seja possível utilizar a metodologia de estimação de Máxima Verossimilhança (ML), é necessário da suposição de alguma distribuição de probabilidade. Dessa maneira, sob condições de regularidade, o estimador de ML é assintoticamente normalmente distribuído, não viesado e totalmente eficiente, com processo de estimação bastante simples quando os dados são gerados por um modelo Gaussiano.

À vista disso, tal como estabelecido por Uribe-Opazo, Borssoi e Galea (2012), aqui vai-se considerar um modelo Gaussiano com uma caracterização linear para a média, isto é, dado um conjunto $s_1, s_2, \dots, s_n$ de localizações distintas e seja $Z(s_1), Z(s_2), \dots, Z(s_n)$ o vetor de observações, então, a tendência espacial é $\mu(s) = \mathbf{X}\beta$ e covariância $\Sigma = \phi_2 \mathbf{R}(\phi_3) + \phi_1 \mathbf{I}$, em que $\mathbf{X}$ é uma matriz de covariáveis, $n \times p$, β é um vetor de parâmetros desconhecidos de ordem $p \times 1$ associados aos efeitos fixos, $\phi_1 \geq 0, \phi_2 \geq 0$ e $\phi_3 > 0$ são os parâmetros que definem a estrutura de dependência espacial, $\mathbf{I}$ é uma matriz identidade de ordem $n \times n$ e $\mathbf{R}(\phi_3)$ é uma matriz de variância e covariâncias de dimensão $n \times n$, com o (i, j) -ésimo elemento igual a $C(h_{ij})$, aqui definida pela estrutura de covariância GW, $C(\|\mathbf{h}\|) = \phi_2 \varphi_{\mu, \kappa}(\|\mathbf{h}\|/\phi_3)$.

Diante disso, supondo que os dados seguem uma distribuição Gaussiana multivariada, obtém-se a função de verossimilhança

$$L(\beta, \phi) = \frac{1}{|\Sigma|^{1/2} \sqrt{(2\pi)^n}} \exp \left[-\frac{(\mathbf{Z} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{Z} - \mathbf{X}\beta)}{2} \right], \quad (2.36)$$

então, aplicando o logaritmo natural em (2.36), obtém-se o logaritmo da função verossimilhança:

$$l(\beta, \phi) = -\frac{1}{2} \log(|\Sigma|) - \frac{n}{2} \log(2\pi) - \frac{1}{2} (\mathbf{Z} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{Z} - \mathbf{X}\beta), \quad (2.37)$$

em que $\phi = (\phi_1, \phi_2, \phi_3)^\top$. As estimativas para o vetor de parâmetros $\theta = (\beta^\top, \phi^\top)^\top$ são obtidas maximizando a equação (2.37), o que em geral produz um sistemas de equações que não possui forma fechada. Diante disso, o vetor de estimativas de máxima verossimilhança pode ser obtido com recursos de técnicas numéricas.

Ademais, Zhang (2004) mostra que para o caso da família Matérn, quando o parâmetro de suavidade (ν) é conhecido e fixo, os parâmetros, variância e escala, não podem ser estimados de forma consistente. Mostrando que este entrave é superado definindo uma relação entre os parâmetros, definindo assim um novo parâmetro nomeado de micro.

Diante disso, Bevilacqua et al. (2019, p.10) utiliza destas ideias para avaliar propriedades assintóticas do estimador de ML do parâmetro micro associado à classe GW, logrando de forte consistência e distribuição assintótica da estimativa deste parâmetro, assumindo que o parâmetro de suporte, aqui nomeado $\phi_3 > 0$ seja fixo e conhecido. Nesta conjuntura, e sob os critérios de equivalência de medidas Gaussianas para a família GW, os autores indica e demonstra algumas propriedades assintóticas para esta classe de correlação, garantindo que os estimadores de ML são obtidos de forma consistente quando se utiliza o parâmetro micro $\tau = \frac{\phi_2}{\phi_3^{2\kappa+1}}$ para GW.

À vista disso, considera-se uma parametrização para estrutura da matriz de covariância espacial, com $\phi_3 > 0$ fixo, tal como em Anderson et al. (1973), Uribe-Opazo, Borssoi e Galea (2012), isto é, concentra-se em uma matriz de covariância com forma linear escrevendo como em (2.38).

$$\mathbf{V} = \phi_1 \mathbf{I}_n + \tau \mathbf{R}(\phi_3), \quad (2.38)$$

então, com esta parametrização fica definido um novo vetor de parâmetros $\theta' = (\phi_1, \tau)$ e $\mathbf{Z}$ tem distribuição normal multivariada de média $\mu = \mathbf{X}\beta$ e covariância $\mathbf{V}$, isto é, $\mathbf{Z} \sim N(\mathbf{X}\beta, \mathbf{V})$. Assim, o logaritmo da função verossimilhança é reescrito como em (2.39).

$$l(\beta, \theta') = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\mathbf{V}|) - \frac{1}{2} (\mathbf{Z} - \mathbf{X}\beta)^\top (\mathbf{V})^{-1} (\mathbf{Z} - \mathbf{X}\beta). \quad (2.39)$$

Posto isso, o próximo passo é efetivar a estimação dos parâmetros dos efeitos fixos β , e da estrutura de dependência espacial θ' . Para tanto, primeiramente obtém-se uma expressão válida para derivada de (2.39) com relação ao parâmetro β , determinando sua forma fechada para que assim seja possível determinar o logaritmo da função verossimilhança concentrada apenas nos parâmetros ϕ_1 e τ . Esta, deve ser maximizada numericamente. Então,

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}, \phi_1, \tau)}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{V})^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \right\} \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2} [\mathbf{Z}^\top (\mathbf{V})^{-1} \mathbf{Z} - \mathbf{Z}^\top (\mathbf{V})^{-1} \mathbf{X}\boldsymbol{\beta} \right. \\
&\quad \left. - \boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{V})^{-1} \mathbf{Z} + \boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{V})^{-1} \mathbf{X}\boldsymbol{\beta}] \right\} \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2} [\mathbf{Z}^\top (\mathbf{V})^{-1} \mathbf{Z} - 2\mathbf{Z}^\top (\mathbf{V})^{-1} \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{V})^{-1} \mathbf{X}\boldsymbol{\beta}] \right\} \\
&= -\frac{1}{2} (-2(\mathbf{Z}^\top (\mathbf{V})^{-1} \mathbf{X})^\top + 2\mathbf{X}^\top (\mathbf{V})^{-1} \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Z} - \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\boldsymbol{\beta},
\end{aligned} \tag{2.40}$$

igualando a expressão em (2.40) a zero, obtém-se:

$$\begin{aligned}
\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Z} - \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\boldsymbol{\beta} &= 0 \\
\Rightarrow \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Z} \\
\Rightarrow \hat{\boldsymbol{\beta}}(\mathbf{V}) &= (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Z},
\end{aligned} \tag{2.41}$$

agora, substituindo (2.41) em (2.39) obtém-se o logaritmo da função verossimilhança concentrada nos parâmetros ϕ_1 e τ ,

$$l(\phi_1, \tau) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\mathbf{V}|) - \frac{1}{2} (\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{V}))^\top \mathbf{V}^{-1} (\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{V})). \tag{2.42}$$

Nesse caso, os estimadores de máxima verossimilhança são determinados maximizando numericamente a função (2.42) e em seguida substituindo nas expressões de forma fechada para se determinar as estimativas do vetor de parâmetro $\boldsymbol{\beta}$.

2.4.3 Erro padrão assintótico

Consoante a Cordeiro e Neto (2004), a inversa da matriz de informação de informação de Fisher $\mathbf{K}$, representa a estrutura de covariância assintótica das estimativas de máxima verossimilhança. Nesse sentido, os erros padrões assintóticos podem ser computados a partir da mesma.

Dessa maneira, sob uma reparametrização adequada para matriz escala, dada por (2.38), os elementos da matriz de informação de Fisher, aqui denotada por $\mathbf{K}(\boldsymbol{\beta}, \boldsymbol{\theta}')$, para o logaritmo da função verossimilhança concentrada (2.39), são calculados como:

$$E \left(-\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} \right) = E (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}) = \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X},$$

$$\begin{aligned}
E\left(-\frac{\partial^2 l}{\partial \theta'_j \partial \theta'_i}\right) &= \frac{1}{2}E\left[\text{tr}\left(\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial \theta'_i \partial \theta'_j}\right)\right] \\
&- \frac{1}{2}E\left[\text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_j}\right)\right] \\
&= \frac{1}{2}\text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_j}\right),
\end{aligned}$$

e

$$\begin{aligned}
E\left(-\frac{\partial^2 l}{\partial \beta \partial \theta'_i}\right) &= E\left(-\frac{\partial^2 l}{\partial \theta'_i \partial \beta}\right) \\
&= E\left[-\left(\mathbf{X}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\mathbf{X}\beta - \mathbf{X}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\mathbf{Z}\right)\right] \\
&= -\mathbf{X}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\mathbf{X}\beta + \mathbf{X}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}E(\mathbf{Z}) \\
&= -\mathbf{X}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\mathbf{X}\beta + \mathbf{X}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\mathbf{X}\beta \\
&= \mathbf{0}
\end{aligned} \tag{2.43}$$

As derivadas da matriz $\mathbf{V}$ para cada caso da função de covariância de Wendland encontra-se no Apêndice C. Posto isto, a matriz de informação de Fisher, para este caso, pode ser escrita como em (2.44).

$$\mathbf{K}(\beta, \theta') = \begin{pmatrix} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_i}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta'_j}\right) \end{pmatrix}. \tag{2.44}$$

Note que, em particular, aqui, diante da parametrização adotada o vetor θ' é instituído dos parâmetros ϕ_1 e τ . Então, neste caso, o bloco correspondente ao segundo elemento da diagonal principal da matriz (2.44) é uma matriz de ordem 2×2 . Logo, a partir dos parâmetros estimados por Máxima Verossimilhança, na Seção que segue realça-se uma valorosa etapa em pesquisas geoestatística a partir da descrição e análise da variabilidade espacial de uma variável de interesse.

Isso posto, no que tange o contexto geoestatístico, segundo Borssoi, Uribe-Opazo e Galea (2011a), a qualidade de mapas depende de boas inferências sobre o modelo ajustado, de modo que na presença de pontos discrepantes, o processo de modelagem deve ser feito com precauções. E nesta conjuntura, os autores utilizam das técnicas de influência local em modelos espaciais lineares Gaussianos. A vista disso, na seção que segue, 2.5, é feita uma síntese desta técnica de diagnóstico para em seguida empregar-se a metodologia no processo de análise de pontos influentes e/ou outliers em modelos espaciais Gaussianos com estrutura de covariância de Wendland.

2.5 DIAGNÓSTICO DE INFLUÊNCIA

Tal como em modelos lineares, modelos lineares mistos, modelos lineares generalizados - MLG's, etc, os modelos espaciais precisam ser validados, no sentido de verificar se as suposições envolvidas são satisfeitas e se o modelo é sensível a pequenas perturbações. Esta metodologia é conhecida na literatura como análise de diagnóstico de influência local.

No que se refere a “análise de diagnóstico de influência local”, esta é uma técnica que consiste em avaliar a qualidade do ajuste à pequenas perturbações nos dados ou no modelo proposto, ou seja, detectar a presença de observações influentes, atípicas e/ou outliers. Neste cenário, é possível analisar o impacto, sobre o aspecto individual, das observações no processo de estimação mediante a exclusão das observações, sendo denominada análise de influência global. Contudo, esta técnica apresenta sérias deficiências devido ao fato da exclusão de todas as informações a partir de uma observação, dificultando inferir sobre a influência da observação acerca do aspecto do modelo. Em contrapartida, é possível desviar-se desse problema analisando o modelo de sensibilidade sob pequenas perturbações. Esta técnica é conhecida como análise de influência local. Nesta conjuntura, Cook (1986) propõe uma medida dessa sensibilidade sob pequenas perturbações nos dados ou no modelo por meio do afastamento da verossimilhança. Desde então, diversos trabalhos tem sido desenvolvidos, dos quais podemos mencionar alguns dos que foram consultados para o desenvolvimento deste trabalho: Galea-Rojas, Bolfarine e Castro (2002) descreveram a metodologia de influência local em aplicações de modelos de calibração comparativa. Os autores Díaz-García, Rojas e Leiva-Sánchez (2003) descreveram métodos diagnósticos para modelos elípticos de regressão multivariada e discutem sobre a técnica de influência local. Castro et al. (2006) descreveram os métodos de influência local para avaliar os efeitos de pequenas perturbações nos dados em estatísticas de teste. Borssoi, Uribe-Opazo e Rojas (2009) descrevem técnicas de influência local com aplicações em geoestatística e análise de dados de agricultura. Assumpção, Opazo e Galea (2011) descreveram métodos de análise de influência local para variável regionalizada com distribuição t-Student multivariada em comparação com normal multivariada. Posto isto, na Subseção 2.5.1 é realizado um compêndio desta metodologia.

2.5.1 Influência Local

Esta metodologia, proposta inicialmente por Cook (1986), apresenta como principal característica utilizar-se das ideias da geometria diferencial para definir uma medida de diagnóstico por meio do afastamento das verossimilhanças, obtidas por um modelo, através de pequenas perturbações no modelo ou diretamente nos dados. Em oposição a técnica de influência global, esta metodologia não exige a delação de observações e nos permite avaliar a influencia conjunta de todos os pontos.

Ao longo destes poucos mais de 34 anos do advento das técnicas de influência local, o processo de perturbação apresenta-se sobre diversas ramificações, podendo ser aplicado na matriz de covariâncias, nos parâmetros da variável resposta, dentre outros. No entanto, na proposta original de Cook (1986), a perturbação é aplicada no afastamento da verossimilhança, isto é, “R. Dennis Cook” sugere investigar o comportamento da função (2.45).

$$\mathcal{LD}(\omega) = 2 \left[l(\hat{\theta}) - l(\hat{\theta}|\omega) \right], \quad (2.45)$$

em que $l(\hat{\theta}|\omega)$ denota o logaritmo da função de verossimilhança perturbada por um vetor de perturbação $\omega = (\omega_1, \dots, \omega_n)^\top \in \Omega \subset \mathbb{R}^n$, $\hat{\theta}$ e $\hat{\theta}_\omega$ são as estimativas de máxima verossimilhança dos modelos não perturbado e perturbado, respectivamente. Ademais, é assumido que existe $\omega_0 \in \Omega$, tal que $l(\hat{\theta}) = l(\hat{\theta}|\omega_0)$, $\forall \hat{\theta}$ no espaço paramétrico, ou seja, ω_0 denota um vetor de não perturbação, e $\mathcal{LD}(\omega) \geq 0$.

Neste contexto, tal qual Paula (2013), a técnica de afastamento pela verossimilhança procura selecionar uma direção unitária $\mathbf{l}$, $|\mathbf{l}| = 1$, e a partir de então analisar o gráfico de $\mathcal{LD}(\omega_0 + c\mathbf{l})$, contra c em que $c \in \mathbb{R}$. Em particular, veja que $\mathcal{LD}(\omega_0) = 0$, então, $l = \mathcal{LD}(\omega_0 + c\mathbf{l})$ tem um mínimo local em $c = 0$. Posto isto, a ideia de Cook (1986), para mensurar a sensibilidade do afastamento da verossimilhança é estudar o comportamento do gráfico de influência na vizinhança de ω_0 , utilizando a curvatura normal, C_1 , ao redor de ω_0 na direção de $\mathbf{l}$. Segundo Paula (2013), a curvatura normal pode ser interpretada como sendo o inverso do raio do melhor círculo ajustado em $c = 0$, sugerindo considerar a direção máxima, $\mathbf{l}_{max}$, que corresponde a maior curvatura, denotada por $C_{1_{max}}$.

O cálculo da curvatura normal requer o uso de ferramentas e conceitos da Geometria Diferencial, e Cook (1986), mostrou que esta, pode ser calculada através de sua equação equação 16, p.8, dada por

$$C_1(\theta) = 2 \left| \mathbf{l}^\top \Delta^\top \ddot{L}_{\hat{\theta}\hat{\theta}}^{-1} \Delta \mathbf{l} \right|, \quad (2.46)$$

em que $-\ddot{L}_{\hat{\theta}\hat{\theta}} = \frac{\partial^2 l(\theta)}{\partial \theta^\top \partial \theta} \Big|_{\theta=\hat{\theta}}$ é a matriz de informação de Fisher e $\Delta = \frac{\partial^2 l(\theta|\omega)}{\partial \theta^\top \partial \omega}$, avaliado em $\omega = \omega_0$ e $\theta = \hat{\theta}$.

De acordo com Paula (2013), o valor máximo de $\mathbf{I}^\top \mathbf{B} \mathbf{I}$, em que $\mathbf{B} = \Delta^\top (-\ddot{L}_{\hat{\theta}\hat{\theta}})^{-1} \Delta$, corresponde ao maior autovalor da matriz $\mathbf{B}$, de tal modo que, a máxima curvatura, $\mathbf{C}_{\mathbf{I}_{max}}$, corresponda ao maior autovalor de $\mathbf{B}$ com $\mathbf{I}_{max}$ o maior autovetor correspondente, permitindo, assim, avaliar o comportamento da curvatura relativa a superfície de influência.

Por conseguinte, a i -ésima componente do vetor $\mathbf{I}_{max}$ indica a influência da i -ésima observação. Então, por estas ideias desenvolvidas por Cook (1986) e exploradas em Paula (2013), se a i -ésima componente de $\mathbf{I}_{max}$ for demasiadamente grande, com relação as demais, então a i -ésima observação correspondente é localmente influente. Neste contexto, Cook (1986) avalia a influência local de $\mathbf{C}_{\mathbf{I}_{max}}$ e $\mathbf{I}_{max}$, utilizando um esquema de casos ponderados em uma perturbação aditiva na variável exploratória em um modelo de regressão linear.

Desta feita, considere o processo estocástico estacionário $\{Z(\mathbf{s}) : \mathbf{s} \in \mathbf{S} \subset \mathbb{R}^d\}$, em que $\mathbf{Z}$ assume distribuição normal multivariada de média $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ e covariância $\boldsymbol{\Sigma} = \phi_1 \mathbf{I}_n + \tau \mathbf{R}(\phi_3)$, sendo $\mathbf{R}(\phi_3)$ uma matriz de covariâncias pertencente a família de Wendland e $\tau = \frac{\phi_2}{\phi_3^{1+2\kappa}}$ o micro parâmetro definido na Equação (2.38). Neste ponto, por simplicidade, nomeia-se o vetor $\boldsymbol{\phi}' = (\phi_1, \tau)$. Considere, ainda, o logaritmo da função de verossimilhança para o modelo não perturbado dada pela expressão (2.39), e a matriz de informação de observada $\mathcal{I}(\boldsymbol{\theta}) = -\ddot{L}(\boldsymbol{\theta})$.

Diante disso, para se iniciar o processo de perturbação na variável resposta, tal qual sugerido em De Bastiani et al. (2015), considere o processo estocástico $\{\mathbf{Z}_\omega(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\omega + \epsilon : \mathbf{s} \in \mathbf{S} \subset \mathbb{R}^d, \omega \in \Omega\}$, em que Ω é um espaço de perturbação e $\mathbf{A}$ é uma matriz $n \times n$ que não depende do vetor $\boldsymbol{\beta}$ ou de ω . Desta forma, para selecionar a matriz $\mathbf{A}$ que de fato produza uma perturbação adequada ao modelo Zhu et al. (2007) estabelece uma metodologia que dá informações de como selecionar uma perturbação apropriada. Então, para desenvolver-se a escolha de uma perturbação adequada para o modelo adotado neste trabalho faz-se necessário emitir a definição 2.5.1.

(ZHU et al., 2007) O produto interno dos operadores ∂_i, ∂_j é

$$g_{ij}(\omega) = \langle \partial_i, \partial_j \rangle = E_\omega [\partial_i l(\omega|\mathbf{Z}, \boldsymbol{\theta}) \partial_j l(\omega|\mathbf{Z}, \boldsymbol{\theta})] \quad (2.47)$$

em que E_ω denota o valor esperado com relação a $f(\mathbf{Z}|\boldsymbol{\theta}, \omega)$. As quantidades n^2 e $g_{ij}(\omega)$, $i, j = 1, \dots, n$ formam um tensor métrico.

A matriz métrica $G(\omega) = (g_{ij}(\omega))$, na Equação (2.47), definida desta forma é uma matriz de informação de Fisher esperada em relação ao vetor de perturbação ω . Seus elementos

medem a quantidade de perturbação que todos os componentes de um vetor de perturbação ω contribuem para um modelo estatístico, (ZHU et al., 2007). À vista disso, o autor estabelece que uma perturbação é apropriada se satisfaz $G(\omega_0) = c\mathbf{I}_n$ com c uma constante positiva.

Posto isto, aqui, baseia-se nestas ideias para mostrar qual perturbação é adequada para o esquema de perturbação na variável resposta em modelos espaciais gaussianos sob uma matriz de covariância do tipo GW. Então, o logaritmo da função de verossimilhança perturbada por um vetor $\omega = (\omega_1, \dots, \omega_n)^\top$ é dada pela Equação (2.48).

$$l(\theta|\omega) = -\frac{1}{2} \log(|\Sigma|) - \frac{n}{2} \log(2\pi) - \frac{1}{2}(\mathbf{Z}\omega - \mathbf{X}\beta)^\top \Sigma^{-1}(\mathbf{Z}\omega - \mathbf{X}\beta), \quad (2.48)$$

em que $\mathbf{Z}\omega$ escrito desta forma é conhecido na literatura como perturbação aditiva na variável resposta. Ademais, para pontos não perturbados, isto é, assumindo $\omega_0 = (\omega_1, \dots, \omega_n)^\top = (0, \dots, 0)^\top$, com $l(\hat{\theta}) = l(\hat{\theta}|\omega_0)$, avalia-se o comportamento local de (2.45) em torno de ω_0 a partir da curvatura normal (2.46) avaliada em ω_0 na direção de um vetor unitário $\mathbf{1}$.

Consequentemente, agora, é preciso determinar as expressões correspondentes aos valores da matriz de informação observada para o logaritmo da função de verossimilhança perturbada. Para tanto, alicerçar-se na teoria construída na seção 2.4.2. Então,

$$\begin{aligned} \frac{\partial l(\theta|\omega)}{\partial \beta} &= \mathbf{X}^\top \Sigma^{-1} \mathbf{Z}\omega - \mathbf{X}^\top \Sigma^{-1} \mathbf{X}\beta \\ &= \mathbf{X}^\top \Sigma^{-1} \mathbf{Z} + \mathbf{X}^\top \Sigma^{-1} \mathbf{A}\omega - \mathbf{X}^\top \Sigma^{-1} \mathbf{X}\beta \\ \Rightarrow \frac{\partial^2 l(\theta|\omega)}{\partial \beta \partial \omega^\top} &= \mathbf{X}^\top \Sigma^{-1} \mathbf{A} \end{aligned}$$

e

$$\frac{\partial l(\theta|\omega)}{\partial \phi'_i} = -\frac{1}{2} \text{tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_i} \right) + \frac{1}{2} (\mathbf{Z}\omega - \mathbf{X}\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_i} \Sigma^{-1} (\mathbf{Z}\omega - \mathbf{X}\beta),$$

então,

$$\begin{aligned} \frac{\partial^2 l(\theta|\omega)}{\partial \phi'_i \partial \omega^\top} &= 2 \cdot \frac{1}{2} \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_i} \Sigma^{-1} (\mathbf{Z}\omega - \mathbf{X}\beta) \times \frac{\partial (\mathbf{Z}\omega - \mathbf{X}\beta)^\top}{\partial \omega^\top} \\ &= (\mathbf{Z} + \mathbf{A}\omega - \mathbf{X}\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_i} \Sigma^{-1} \\ &= (\mathbf{Z} + \mathbf{A}\omega - \mathbf{X}\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_i} \Sigma^{-1} \mathbf{A}, \end{aligned} \quad (2.49)$$

em que $i = 1, 2$, tendo em vista que, para este caso, o vetor de parâmetros utilizado é o definido na reparametrização (2.38). Desta forma, avaliando (2.49) em $\omega = \omega_0$ se consegue os elementos da matriz Δ , com $\Delta = \left(\Delta_\beta^\top, \Delta_{\phi'}^\top \right)^\top$, em que $\Delta_{\phi'} = \left(\Delta_1^\top, \dots, \Delta_q^\top \right)^\top$ e $\Delta_i = (\mathbf{Z} - \mathbf{X}\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_i} \Sigma^{-1} \mathbf{A}$, com $i = 1, \dots, q$. Em nosso caso $q = 2$. À vista disso, tem-se que

$$\Delta = \left(\left(\mathbf{X}^\top \Sigma^{-1} \mathbf{A} \right)^\top, \left((\mathbf{Z} - \mathbf{X}\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_1} \Sigma^{-1} \mathbf{A} \right)^\top, \left((\mathbf{Z} - \mathbf{X}\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi'_2} \Sigma^{-1} \mathbf{A} \right)^\top \right)^\top.$$

Agora, da definição 2.5.1 considere a matriz $G(\omega) = E_\omega [\mathbf{U}(\omega)\mathbf{U}(\omega)^\top]$ em que $\mathbf{U}(\omega)$ é o vetor escore para o logaritmo da função verossimilhança perturbada dada por:

$$\begin{aligned} \mathbf{U}(\omega) &= -2\frac{1}{2}\Sigma^{-1}(\mathbf{Z}\omega - \mathbf{X}\beta) \frac{\partial}{\partial \omega} (\mathbf{Z}\omega - \mathbf{X}\beta)^\top \\ &= -\mathbf{A}^\top \Sigma^{-1}(\mathbf{Z}\omega - \mathbf{X}\beta), \end{aligned}$$

então,

$$\begin{aligned} G(\omega_0) &= E_{\omega_0} [\mathbf{A}^\top \Sigma^{-1}(\mathbf{Z}\omega_0 - \mathbf{X}\beta)(\mathbf{Z}\omega_0 - \mathbf{X}\beta)^\top \Sigma^{-1} \mathbf{A}] \\ &= \mathbf{A}^\top \Sigma^{-1} \mathbf{A}. \end{aligned} \quad (2.50)$$

Note que em (2.50) usa-se o fato de $E_{\omega_0}[(\mathbf{Z}\omega_0 - \mathbf{X}\beta)(\mathbf{Z}\omega_0 - \mathbf{X}\beta)^\top] = \Sigma$. Sendo assim, considerando $\mathbf{A} = \Sigma^{1/2}$ e $c = 1$, tem-se:

$$\begin{aligned} G(\omega_0) &= \Sigma^{1/2} \Sigma^{-1} \Sigma^{1/2} \\ &= (\Sigma^{1/2})^{-1} \Sigma^{1/2} = \mathbf{I}_n. \end{aligned}$$

Portanto, no caso particular deste trabalho, uma perturbação adequada é do tipo $\mu(\omega) = \mathbf{X}\beta + \Sigma^{1/2}\omega$ que coincide com o obtido por De Bastiani et al. (2015).

Nesta conjuntura, tal como disponível na literatura existe diversas maneiras de se utilizar estes resultados para se estudar a superfície de influência. Uma destas é analisar o gráfico de $|lmax| \times i$ (ordem das observações) para verificar, segundo Cook (1986), qual tipo de perturbação exerce maior influência no afastamento da verossimilhança em uma vizinhança de w_0 . Esta análise gráfica pode indicar quais das observações exercem maior influência sobre $\mathcal{LD}(\mathbf{w})$ em uma vizinhança de $\mathbf{w}_0$. Estas, por sua vez, podem ser agentes de mudanças substanciais nas estimativas dos parâmetros sob pequenas perturbações nos dados ou mesmo no modelo. Ademais, segundo Paula (1993) quando a diferença entre o primeiro e segundo autovalor for pequena, pode ser informativo que se olhe os componentes do segundo autovetor, pois, neste caso, é provável que se destaque algum tipo de influência particular das observações nas estimativas.

Ainda, aqui está sendo utilizado como vetores unitários (1) os vetores usuais da base canônica do $\mathbb{R}^n$, e_i . Neste caso, conforme Uribe-Opazo, Borssoi e Galea (2012) a curvatura

normal, dita como influência local da i -ésima observação, é obtida por $C_i = 2|b_{ii}|$ em que b_{ii} indica o i -ésimo elemento da diagonal principal da matriz $\mathbf{B} = \Delta^\top (-\ddot{L}_{\hat{\theta}\hat{\theta}})^{-1} \Delta$. Deste modo, a análise gráfica de $C_i \times i$ (ordem das observações) é indicada para avaliar a existência de observações influentes.

No entanto, na interpretação dos resultados precedentes surge um problema de determinar qual critério utilizar para indicar quando a medida encontrada é grande o suficiente para que a observação seja considerada influente. À vista disso, diferentes critérios são propostos para avaliar a influência local das observações. Borssoi, Uribe-Opazo e Galea (2011b) sugere considerar a i -ésima observação como influente se seu valor for maior que duas vezes a média das medidas $C_i = 2|b_{ii}|$. Cook (1986) considerou em duas unidades uma referência para C_{max} em uma amostra aleatória simples. Poon e Poon (1999), no contexto dos modelos de regressão linear, idealizaram níveis de referência a partir das medidas de influência local através de uma curvatura normal conformal e C_{max} . Sendo a curvatura normal conformal é dada $\mathbf{F}_l = \frac{C_l}{tr(2\mathbf{B})}$ garantindo a propriedade $0 \leq \mathbf{F}_l \leq 1$. Neste caso, definindo a medida $F_i = \frac{2|b_{ii}|}{tr(2\mathbf{B})}$ na direção do i -ésimo vetor canônico do $\mathbb{R}^n$, Zhu e Lee (2001) definiram a i -ésima observação como potencialmente influente se $F_i > \frac{\sum_{i=1}^n F_i}{n} + 2sd(\mathbf{F})$, em que $sd(\mathbf{F})$ denota o desvio padrão para $F_1, \dots, F_n$. Esta medida sendo utilizada posteriormente por De Bastiani et al. (2015) no contexto dos modelos elípticos.

Note que a abordagem precedente diz respeito a análise de influência local em um contexto global. No tocante à abordagem individual, Cook (1986) propôs comparar os grandes componentes dos resíduos do modelo de regressão linear com os maiores valores de l_{max} (autovetor correspondente o maior autovalor da matriz $\mathbf{B}$). (ZHANG; KING, 2005) construíram linhas de referência, utilizando simulações de Monte Carlo, para investigar a significância das medidas de influência local. Zhu e Zhang (2004) por meio de uma perturbação ponderada mostrou que a medida de influência local, a distância de Cook e a distância de probabilidades são assintoticamente equivalentes.

Diante disso, neste trabalho é proposto a classificação das observações como influentes construindo uma linha de corte por simulações *Bootstrap* como na Subseção 2.5.2.

2.5.2 Níveis de referência para classificar observações potencialmente influentes

Uma etapa importante na análise do diagnóstico de influência é determinar os critérios de validação quando a medida encontrada é grande o suficiente para que a observação correspondente seja considerada potencialmente influente.

Diante disso, recentemente Martin e Roberts (2010) adaptou a teoria Jackknife-after-Bootstrap desenvolvida por Efron (1992) para obter pontos de influência a partir da distância de Cook de (COOK, 1977).

A ideia da técnica jackknife-after-bootstrap consiste em um processo de validação cruzada, em que cada subgrupo é obtido excluindo a i -ésima observação, de forma que a distribuição bootstrap da medida de diagnóstico em estudo, aproximada por reamostragem, não seja contaminada por este ponto. De acordo com Davison e Hinkley (1997, p. 114) a probabilidade de que o ponto z_i não apareça em uma amostra de bootstrap de tamanho n é $(1 - n^{-1})^n \approx e^{-1}$, então o número de simulações B_j que não incluem z_i é aproximadamente igual a $B_j e^{-1} = 0,368B_j$, em que o índice j varia nos casos de reamostragem.

Assim, este trabalho se propõe a classificar as observações como potencialmente influentes por meio da construção de um ponto de corte (ou nível de referência) por meio de simulações Bootstrap. O algoritmo proposto por Martin e Roberts (2010) é adaptado para classificar observações influentes a partir da medida de influência local (C_i). Essa motivação se justifica pelo fato dessa metodologia ainda não ter sido abordada para esse tipo de análise. Assim, os níveis de referência são construídos usando o algoritmo 1.

Algorithm 1 Ponto de corte por Bootstrap

Passo 1. Seja C_i a estatística de diagnóstico em estudo. Calcule os valores C_i , $i = 1, \dots, n$ do modelo proposto;

Passo 2. Calcule B amostras Bootstrap com reposição do conjunto de dados originais;

Passo 3. Para cada ponto z_i , obtenha B re-amostras, e determine o subconjunto das amostras que não contêm o ponto z_i , aproximadamente $B_j e^{-1}$, em que o índice j varia nos casos de reamostragem. Em seguida, calcule os valores C_i , $i = 1, 2, \dots, n$ para este subconjunto;

Passo 4. Armazene os $nB_j e^{-1}$ valores de C_i em um único vetor;

Passo 5. O quantil 95% da distribuição gerada é usado como um ponto de corte para medir a influência local C_i . Se o valor de C_i for maior que o quantil sugerido, a observação correspondente será marcada como influente. Caso contrário, volte para o Passo 1.

De acordo com Martin e Roberts (2010), a ideia no passo 3 é gerar uma distribuição bootstrap nula da medida de influência sob a hipótese de que a i -ésima observação não é influente. O autor esclarece que como a i -ésima observação não está presente em nenhuma das re-amostragens a partir das quais essa a distribuição bootstrap é gerada, ela não pode exercer influência e, portanto, a distribuição gerada está livre da influência da mesma.

Ademais, a abordagem realizada nesta seção é feita utilizando um esquema de perturbação aditiva na variável resposta, sendo a influência local analisada por meio do afastamento das verossimilhanças. Para outras técnicas de influência o leitor interessado, por exemplo pode consultar Uribe-Opazo, Borssoi e Galea (2012), o qual desenvolve expressões para análise de influência local na matriz de correlação dos dados, no preditor linear e nos valores preditos.

2.6 SIMULAÇÃO

Diante da necessidade de avaliar o desempenho da análise de influência local sob o esquema de perturbação proposto, aqui, será feito um sucinto estudo de simulação de Monte Carlo em 3 cenários, com estrutura espacial conhecida e, 1000 réplicas de amostras com tamanho $n = 74$. A distribuição adotada provém de um processo espacial gaussiano. O modelo de covariância usado corresponde ao pertencente a GW com parâmetro de suavização $\kappa = 3$, (2.11), selecionado pelo critério de validação cruzada mostrado em De Bastiani et al. (2015). O *software* R Core Team (2020) foi utilizado para realizar todos os cálculos, incluindo os resultados obtidos na Seção 2.7.

Aqui, faz-se uso de um esquema de perturbação na resposta proposto por Ortega, Bolfarine e Paula (2003, p.11): $z_{max} = z_{max} + a\sqrt{\|\mathbf{z}\|}$, em que z_{max} denota o valor máximo do vetor $\mathbf{z}$, $\|\mathbf{z}\| = \mathbf{z}^T \mathbf{z}$ para vários valores de a no intervalo $(0, 1]$. Foram considerados os três seguintes cenários **1**. $a = 0,5$, $(\phi_1, \phi_2, \phi_3)^T = (0, 15; 0, 3; 1)^T$ e $\beta_0 = 1$; **2**. $a = 0,75$, $(\phi_1, \phi_2, \phi_3)^T = (0, 15; 0, 3; 1)^T$ e $\beta_0 = 1$; **3**. $a = 1$, $(\phi_1, \phi_2, \phi_3)^T = (0, 15; 0, 3; 1)^T$ e $\beta_0 = 1$. O parâmetro ϕ_3 é considerado fixo e então a reparametrização $\tau_1 = \phi_1$, $\tau_2 = \frac{\phi_2}{\phi_3^{2\kappa+1}}$ é adotada.

Com relação à eficiência da metodologia na identificação da observação de perturbação, as medidas C_i e l_{max} alcançaram eficiência de 100% em todos os cenários para 1000 simulações. Para resumir, considera-se o cenário $a = 1$ e $(\phi_1, \phi_2, \phi_3)^T = (0, 15; 0, 3; 1)$ e $\beta_0 = 1$, seleciona-se uma amostra dentre as 1000 simuladas, então, os resultados obtidos a partir desta amostra são apresentados. Os gráficos de C_i versus ordem das observações, para dados simulados e perturbados são dados na Figura 7. Notou-se a influência da observação 16, observação

perturbada, que confirma a sensibilidade dos estimadores de ML na presença de dados atípicos. A influência de pontos atípicos no estimador ML é vista nos resultados da Tabela 5.

Assim sendo, e diante dos resultados obtidos pela análise de diagnóstico de influência local, é notado que os gráficos $C_i \times i$, Figura 8(b), segundo os critérios de níveis de referência, identificou a observação perturbada como influente. Portanto, o estudo mostrou que o gráfico $C_i \times i$, foi eficaz no processo de identificação das observações sob o esquema de perturbação adotado.

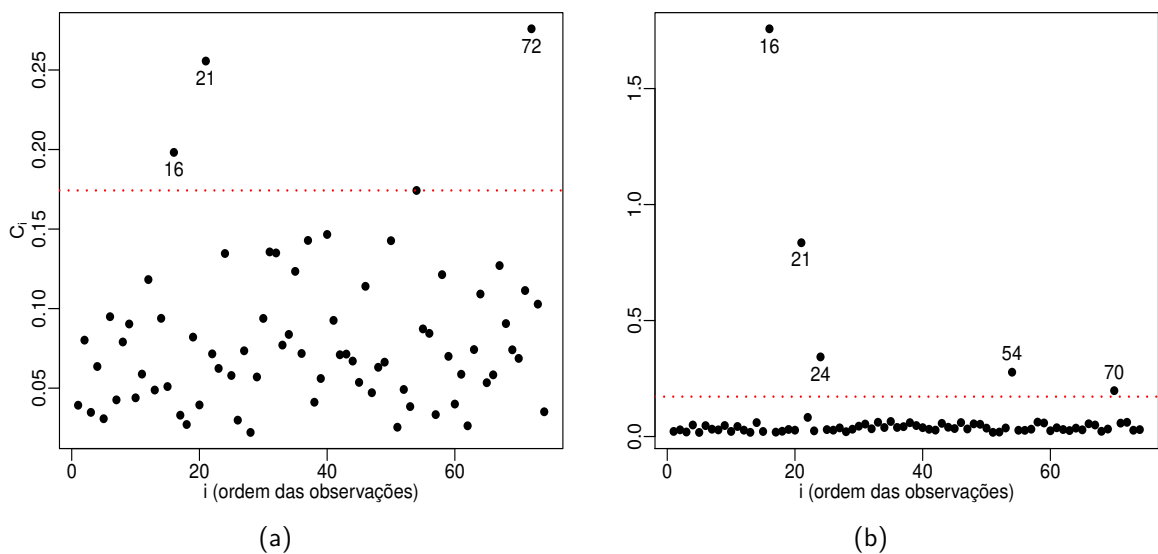
Tabela 5 – Parâmetros estimados por ML para os dados de uma amostra selecionada dentre as 1000 simuladas, considerando a família de covariância Wendland generalizada com parâmetro de suavização $\kappa = 3$ e parâmetros originais $(\tau_1, \tau_2)^T = (0, 15; 0, 3)^T$, $\phi_3 = 1$ e $\beta_0 = 1$.

Dados	Intercepto (Média)	Efeito pepita	Contribuição	AIC	IDE(%)	Var
	$\hat{\beta}_0$	$\hat{\tau}_1$	$\hat{\tau}_2$	— — —	— — —	— — —
Dados Simulados	1,066 (0,195)	0,147 (0,103)	0,244 (0,104)	105,854 — — —	26,287 — — —	0,370 — — —
Dados Perturbados(#16)	1,237 (0,407)	1,294 (0,508)	0,963 (0,502)	255,119 — — —	17,994 — — —	2,025 — — —

Fonte: Elaborada pelo autor (2022).

AIC: Critério de informação de Akaike . IDE: Índice de dependência espacial. Var: Variância.

Figura 7 – Gráficos de influência local para os dados simulados e perturbado. Em (a) é indicado o gráfico da curvatura normal, C_i para os dados simulados, enquanto em (b) a representação é para os dados após a perturbação *versus* ordem das observações.



Fonte: Elaborada pelo autor (2020).

Destarte, na Seção 2.7 é retomada as análises da metodologia de diagnóstico de influência

local no contexto de uma aplicação à dados reais.

2.7 APLICAÇÃO

Com o propósito de ilustrar a aplicação das técnicas geoestatísticas, aqui apresentadas, são realizadas análises em um conjunto de dados reais relacionados a agricultura. No entanto, vale salientar que a proposta aqui não é discutir os pormenores dos dados. Mas sim, verificar o desempenho da metodologia sugerida, a partir de uma função de covariância espacial construídas por funções de Wendland, para quantificar e/ou avaliar o nível da variabilidade espacial do rendimento de grãos de soja. Portanto, como consequência dos resultados obtidos, é possível reconstruir a região contínua do estudo através de mapas, que constituem uma valiosa ferramenta no gerenciamento da variabilidade espacial, sobretudo no que diz respeito a tomada de decisões e desempenho de produção agrícola.

Os dados foram nos fornecidos pelo Laboratório de Estatística Espacial (LEE) da Universidade do Oeste do Paraná (Unioeste), campus Cascavel (PR). A coleta foi realizada em uma região amostral corresponde a uma área comercial de 127,18 *ha*, de produção de grãos, cujas coordenadas geográficas são, aproximadamente, 24,95 *S* e 53,57 *W*, com altitude média de 650 *m*, variável de interesse produtividade ou rendimento da soja (PROD) [*t ha*⁻¹] (tonelada por hectare). O objetivo do experimento foi estudar a variabilidade da produtividade de grãos de soja, no ano agrícola 2016/2017 com 74 observações, em função de atributos do solo: fósforo (*P*) *mg dm*⁻³ (miligrama por decímetro cubico); potássio (*K*) *cmolc dm*⁻³ (Centimol de carga (*cmolc*) de um nutriente por decímetro cubico); Cálcio (*Ca*) *cmolc dm*⁻³ e Magnésio (*Mg*) *cmolc dm*⁻³.

Dito isto, com o objetivo de verificar o comportamento da dispersão dos dados, inicialmente, é realizada uma análise descritiva. As estatísticas descritiva para os dados de produtividade do grão de soja estão na Tabela 6. Percebe-se que os valores da média e mediana estão próximos, então, o que pode ser indício de uma baixa dispersão dos dados em relação a sua posição central. Ainda, verifica-se uma média homogeneidade dos dados vista pelo valor do coeficiente de variação (*Coefficientedevariação(CV)*). Pela análise gráfica do *box-plot* (Figura 9(b)), percebe-se a presença dos pontos discrepantes 1,677, 1,661, 1,518, e 1,797, localizados nas posições amostrais 65, 67, 68 e 69, respectivamente.

Tabela 6 – Estatísticas descritivas para os dados da produtividade da soja (PROD) [$t \text{ ha}^{-1}$] ano 2016/2017 de uma área comercial de 127,18 ha , de produção de grãos, localizada próximo ao município de Cascavel, região Oeste do Paraná (PR)

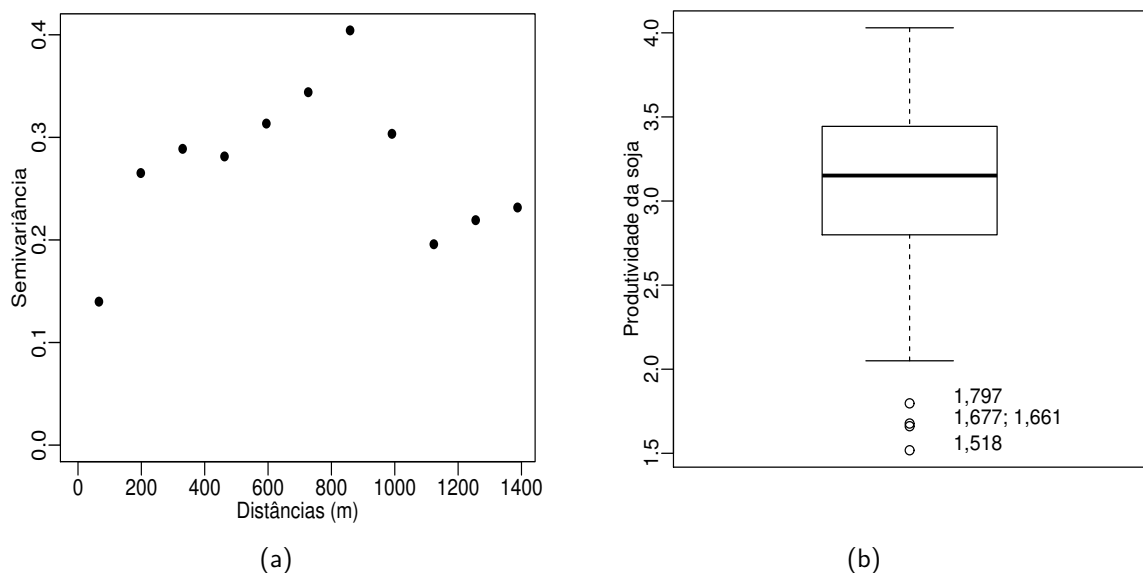
Variável	n	Média	Q_1	Mediana	Q_3	Min.	Max.	SD	Var	CV (%)
Prod	74	3,057	2,799	3,151	3,438	1,518	4,030	0,547	0,299	17,911

Fonte: Elaborada pelo autor (2022).

Prod: Produtividade da soja, n: número de observações, Min.: mínimo, 1st Q_1 : primeiro quartil, Q_3 : terceiro quartil, Max.: máximo, SD: Desvio padrão, CV: coeficiente de variação. Var: Variância.

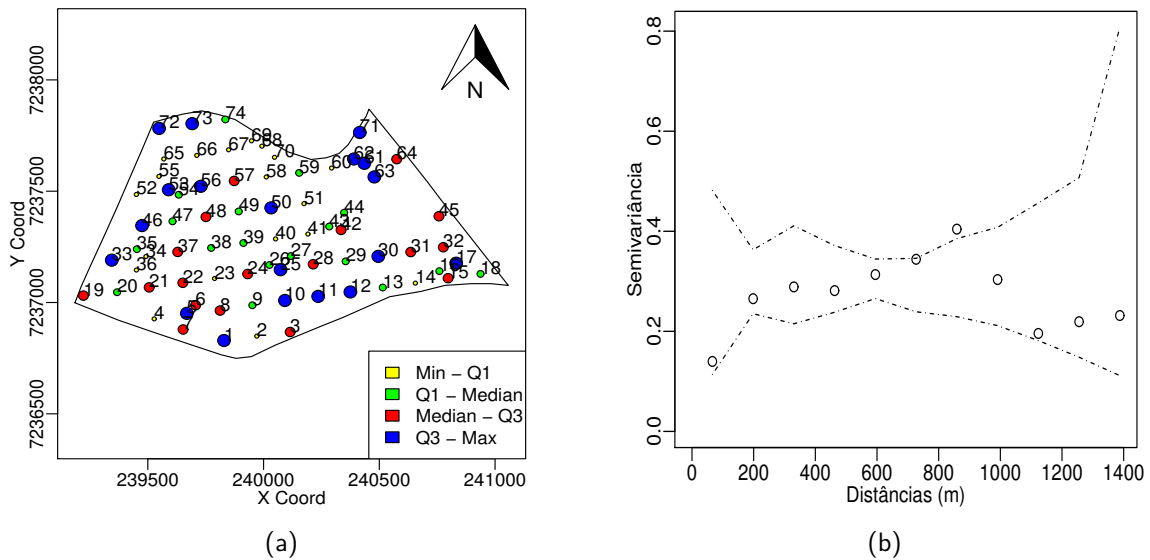
Pelo gráfico *post-plot* (Figura 10(a)), percebe-se que os valores altos estão distribuídos em toda área amostral, assim, verifica-se visualmente que não existe tendência direcional. Logo, é permitido a suposição de estacionariedade intrínseca, sendo, portanto, possível ajustar um modelo de semivariância para explicar a variabilidade espacial do rendimento de soja. A verificação da existência de dependência espacial é realizada pelo gráfico de envelope (Figura 10(b)), em que percebe-se a presença de um ponto fora do envelope, portanto, existe dependência espacial da variável rendimento de soja. Com relação a isotropia, para dados de produtividade da soja, foi verificada no trabalho de Dalposso, Uribe-Opazo e Bastiani (2021).

Figura 8 – Semivariograma experimental, 9(a), boxplot em 9(b) para variável produtividade da soja em ($t \text{ ha}^{-1}$) referente ao ano agrícola 2016/2017.



Fonte: Elaborada pelo autor (2022).

Figura 9 – *Post-plot* da distribuição espacial dos pontos em 10(a) e gráfico de envelope das funções semi-variâncias em 10(b) para variável produtividade da soja em ($t\ ha^{-1}$) referente ao ano agrícola 2016/2017.



Fonte: Elaborada pelo autor (2022).

2.7.1 Modelagem geoestatística

Para modelagem do rendimento médio de soja considerando uma estrutura de dependência espacial, adota-se um processo estocástico Gaussiano $\{Z(s) : s \in \mathbb{R}^2\}$. Desta forma, considera-se um conjunto de realizações $Z(s_1), \dots, Z(s_n)$ conhecidas em diferentes localizações espaciais $s_1, \dots, s_n$. Portanto, admite-se que os dados podem ser mapeados pelo modelo linear espacial

$$Z(s_i) = \mu(s_i) + \epsilon(s_i),$$

em que $\mu(s_i)$ corresponde a parte determinística ou tendência do processo, sendo função das variáveis explicativas. Por outro lado, o termo $\epsilon(s_i)$ denota a componente estocástica, que assume-se ter média zero e variância determinada pela função de covariância $C(s_i, s_i + h) = Cov(\epsilon(s_i), \epsilon(s_i + h))$, sendo h um incremento para localização espacial s_i . Desta maneira, para definir a estrutura de dependência espacial, considera-se o vetor paramétrico $\phi = (\phi_1, \phi_2, \phi_3)$, cujas entradas indicam o efeito pepita, variação estrutural e escala, respectivamente. Sendo assim, considera-se que variável rendimento de soja assume uma distribuição normal multivariada $Z \sim (X\beta, \phi_1 I_n + \phi_2 R(\phi_3))$, em que X e β é uma matriz de ordem $n \times p$ de covariáveis e um vetor paramétrico de ordem $p \times 1$, respectivamente. Enquanto $R(\phi_3)$ denota alguma

matriz de covariância espacial. Aqui, $\mathbf{R}$ são obtidas das famílias de covariâncias Wendland e Matérn.

Contudo, nos trabalhos de Zhang (2004) e Bevilacqua et al. (2019) é mostrado que se o parâmetro de suavização, para as famílias Matérn e Wendland, respectivamente, são fixos e conhecidos, então, as estimativas dos parâmetros não são consistente. Com isso, como alternativa foi sugerido uma reparametrização definindo o parâmetro micro $\tau = \frac{\phi_2}{\phi_3^{2\kappa}}$ (Dalposso, Uribe-Opazo e Bastiani (2021)) e $\tau = \frac{\phi_2}{\phi_3^{2\kappa+1}}$ (Bevilacqua et al. (2019)), para as famílias Matérn e Wendland, respectivamente. Desta maneira, considerando ϕ_3 fixo, fazendo $\phi_1 = \tau_1$ e $\phi_2 = \tau$ então, a matriz de covariância assume a forma linear $\Sigma = \tau_1 \mathbf{I}_n + \tau \mathbf{R}(\phi_3)$.

Desta forma, considerando o uso de covariáveis para o estudo, então, o valor médio para o rendimento da soja na localização espacial $\mathbf{s} \in \mathbb{R}^2$ fica determinado por (2.51).

$$\mu(\mathbf{s}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \quad (2.51)$$

em que x_1, x_2, x_3 e x_4 , denotam as covariáveis fósforo (P), potássio (K), Cálcio (Ca) e Magnésio (Mg), respectivamente, e x_0 é um vetor de um's. Estes, comporão as entradas das linhas da matriz $\mathbf{X}$, enquanto os valores dos β 's comporão o vetor de parâmetros a serem estimados.

Os parâmetros de ajuste foram estimados por Máxima Verossimilhança. Para determinar os parâmetros iniciais utilizou-se o método de estimação dos Mínimos Quadrados Ordinários (OLS). Os detalhes de ambos os métodos foram apresentados na Seção 2.4. O parâmetro que controla o decaimento da função de correlação (ϕ_3) foi considerado fixo iguais a 1,609; 1,609; 1,73, 1,2; 0,5; 0,379; 0,273 e 0,224, nesta ordem para os parâmetros de suavização $\kappa = 0, 1, 2$ e 3 , $\nu = 0,5; 1,5; 2,5$ e $3,5$, para famílias Wendland e Matérn, respectivamente, os quais foram estimados por OLS. Os resultados estão na Tabela 7. Os erros padrões estão indicados entre parênteses.

Observa-se que os parâmetros estimados para a parte determinística β 's, foram semelhantes para todos os casos das funções de correlação Wendland para vários valores de suavização κ . Similarmente, considerando os diferentes valores de suavização da família Matérn, os parâmetros da média obtiveram pouca variação. Na comparação entre as duas famílias notou-se que em todos os casos houve perda de cálcio (Ca), indicando uma deficiência no solo, covariável esta importante para o crescimento da cultura. Com relação aos parâmetros que definem a estrutura de dependência espacial, para a Wendland o efeito pepita foi maior que a variação estrutural, exceto para $\kappa = 2$. Por outro lado, a menos com suavização de $3,5$, para família

Matérn os resultados foram estritamente inverso. Com relação a força da dependência, temos que a força do modelo vista pela medida FM aumenta de acordo com a escolha do parâmetro de suavização. Intuitivamente, em analogia com o índice de dependência definido em Biondi, Myers e Avery (1994), o índice de dependência espacial IDE indica uma moderada dependência espacial para os casos da Wendland. No caso da Matérn o IDE indica forte dependência espacial com $\kappa = 0,5; 1,5; 2,5$ e moderada para $\kappa = 3,5$. Obviamente tais resultados são influenciados pelas estimativa dos parâmetros que dependem dos valores iniciais, então, afirmativas neste sentido devem ser avaliadas com cautela.

Tabela 7 – Parâmetros estimados por método máxima verossimilhança para modelo com covariância espacial de Wendland (W) e Matérn (M) para vários valores do parâmetro de suavização κ e ν , para os dados da variável produtividade da soja em ($t \times ha^{-1}$) do ano agrícola 2016/2017, com $\phi_3 = 1,2$.

κ	$\hat{\beta}_0$ ---	$\hat{\beta}_1$ <i>P</i>	$\hat{\beta}_2$ <i>K</i>	$\hat{\beta}_3$ <i>Ca</i>	$\hat{\beta}_4$ <i>Mg</i>	$\hat{\phi}_1$ ---	$\hat{\phi}_2$ ---	<i>CV</i> ---	<i>AIC</i> ---	<i>IDE</i> (%) (<i>FM</i>)
<i>W</i> – 0	3,480 (0,394)	0,003 (0,006)	-0,395 (0,577)	-0,009 (0,010)	0,067 (0,100)	0,222 (0,079)	0,185 (0,137)	0,372 ---	118,226 ---	19,303 (0,425)
<i>W</i> – 1	3,513 (0,370)	0,002 (0,006)	-0,442 (0,585)	-0,009 (0,010)	0,059 (0,099)	0,250 (0,060)	0,223 (0,203)	0,368 ---	118,723 ---	23,665 (0,502)
<i>W</i> – 2	3,504 (0,367)	0,002 (0,006)	-0,434 (0,585)	-0,009 (0,010)	0,059 (0,099)	0,250 (0,058)	0,767 (0,704)	0,378 ---	118,587 ---	38,525 (0,511)
<i>W</i> – 3	3,402 (0,362)	0,003 (0,006)	-0,388 (0,581)	-0,009 (0,010)	0,069 (0,099)	0,232 (0,053)	0,209 (0,159)	0,367 ---	117,064 ---	17,412 (0,512)
<i>M</i> – 0,5	2,918 (0,330)	0,003 (0,006)	0,056 (0,552)	-0,003 (0,009)	0,108 (0,100)	0,000 (0,091)	0,273 (0,092)	0,377 ---	121,725 ---	31,700 (0,317)
<i>M</i> – 1,5	2,934 (0,320)	0,004 (0,006)	-0,038 (0,550)	-0,003 (0,009)	0,105 (0,100)	0,034 (0,119)	0,239 (0,120)	0,382 ---	120,967 ---	35,827 (0,409)
<i>M</i> – 2,5	2,934 (0,320)	0,004 (0,006)	-0,044 (0,548)	-0,003 (0,009)	0,105 (0,100)	0,052 (0,176)	0,221 (0,136)	0,382 ---	120,787 ---	35,519 (0,439)
<i>M</i> – 3,5	2,881 (0,350)	0,003 (0,006)	0,149 (0,574)	-0,003 (0,010)	0,122 (0,104)	0,147 (0,262)	0,140 (0,120)	0,373 ---	120,480 ---	21,628 (0,444)

Fonte: Elaborada pelo autor (2022).

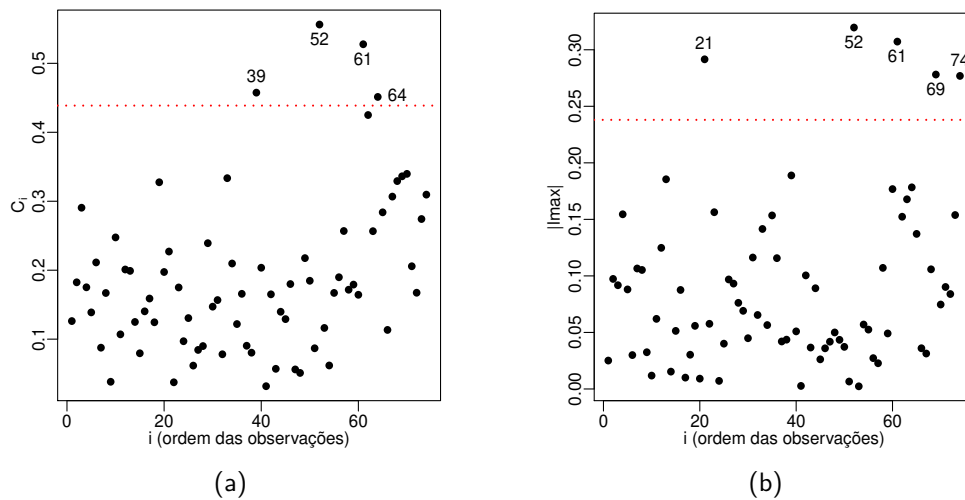
Nota: *CV* = Validação cruzada. *M* - κ : Matérn com suavização κ . *W* - κ : Wendland com suavização κ . (*P*)-fósforo, (*K*)-potássio, (*Ca*)-Cálcio e (*Mg*)-Magnésio.

Desta forma, a escolha da estrutura de covariância para o modelo de ajuste é realizada pelo critério de validação cruzada (*CV*) (detalhes podem ser consultados em De Bastiani (2016)) e *AIC*. Assim, foi selecionado o modelo com estrutura de covariância Wendland cujo parâmetro de suavização é $\kappa = 3$, pois, o mesmo apresentou o menor valor para validação cruzada $CV = 0,367$ e $AIC = 117,064$. Então com o parâmetro de suporte $\phi_3 = 1,2$ tem-se um raio de dependência espacial entre amostras igual a 0,6168 metros.

2.7.2 Influência local

Nesta subseção são aplicadas as técnicas de diagnóstico de influência local, através de gráficos dos coeficientes $(C_i \times i)$ e $|l_{max}|$, com objetivo de identificar os pontos que possam ser influentes na estrutura de dependência espacial da variável produtividade da soja. A medida correspondente aos valores de C_i e l_{max} são denominadas de influência local total e direção máxima da curvatura normal, respectivamente. A metodologia é empregada utilizando um esquema de perturbação do tipo aditivo na resposta $Z\omega = Z + \sqrt{\|Z\|}\omega$ apresentada na Seção 2.5.1, em que ω denota um vetor de perturbação. Para selecionar os níveis de referência, quer dizer, a linha de corte para classificação de observações como influentes, descrita pela linha horizontal pontilhada (Figuras 11(a) e 11(b)), é aplicada a técnica de classificação a partir de intervalos percentílico obtidos por *jackknife-after-bootstrap* definida em 2.5.2.

Figura 10 – Influência local a partir da curvatura normal total $C_i \times$ ordem das observações (11(a)), e direção máxima da curvatura normal $|l_{max}| \times$ ordem das observações (11(b)) para o modelo selecionado de produtividade da soja.



Fonte: Elaborada pelo autor (2022).

Notamos que as observações #39, #52, #61 e #64, correspondendo aos valores de produtividade 2,927; 3,511; 3,543 e 3,400 ($t \text{ ha}^{-1}$) (toneladas por hectare), respectivamente, foram destacadas como potencialmente influentes no gráfico de C_i (Figura 11(a)). No gráfico (Figura 11(b)) destacou-se como potencialmente influentes as observações #21, #52, #61, #69 e #74, correspondendo aos valores de produtividade 3,208; 3,511; 1,797 e 2,835 ($t \text{ ha}^{-1}$) (toneladas por hectare), respectivamente. Dentre estes, a observação #69 também foi destacada no gráfico box-plot (Figura 9(b)).

Com isso, tendo em vista que as observações #52 e #61 foram destacadas em ambos os gráficos, e a observação #69 identificada no box-plot e $|l_{max}|$, então, estes pontos foram retirados do conjunto de dados e realizada a análise novamente. A partir dos resultados obtidos na Tabela 8, verifica-se que observações #52, #61 e #69 exercem pouco influência nas estatísticas descritivas, pois houve pouca variação nos resultados.

Tabela 8 – Estatísticas descritivas para variável produtividade da soja ($\text{Prod}(t \text{ ha}^{-1})$) com e sem a presença das observações potencialmente influentes ($\text{Prod}^*(t \text{ ha}^{-1})$), e quando excluídas individualmente #52, #61 e #69.

Variável	n	Média	Q_1	Mediana	Q_3	Min.	Max.	SD	Var	CV (%)
Prod	74	3,057	2,799	3,151	3,438	1,518	4,030	0,547	0,299	17,911
Prod*	71	3,071	2,817	3,159	3,433	1,518	4,030	0,534	0,285	17,394
#52	73	3,060	2,800	3,159	3,444	1,518	4,030	0,550	0,303	17,986
#61	73	3,050	2,799	3,144	3,421	1,518	4,030	0,548	0,300	17,975
#69	73	3,074	2,800	3,159	3,444	1,518	4,030	0,531	0,282	17,262

Fonte: Elaborada pelo autor (2022).

Prod: Produtividade da soja, n: número de observações, Min.: mínimo, 1st Q_1 : primeiro quartil, Q_3 : terceiro quartil, Max.: máximo, SD: Desvio padrão, CV: coeficiente de variação. Var: Variância.

2.7.2.1 Influência das observações potencialmente influentes nas estimativas dos parâmetros

Fundamentado na análise de influência precedente para o modelo proposto, aqui, expõe-se os resultados e considerações do processo de estimação dos parâmetros, por máxima verossimilhança, para o modelo Wendland com suavização $\kappa = 3$, quando se é excluída simultaneamente e/ou individualmente as observações classificadas como potencialmente influentes.

Com isso, e com vistas na Tabela 9, verifica-se que as estimativas para o parâmetro do efeito pepita (ϕ_1), não houve aumento apenas na exclusão individual da observação #52, em que manteve o mesmo resultado quando dos dados completos. Neste mesmo cenário, o parâmetro da variância estrutural ou contribuição (ϕ_2), na aproximação, é permitido afirmar que também não houve alteração com relação as estimativas com os dados completos. Ademais, as influências significativas na estrutura de dependência espacial, vistas nas alterações das estimativas de ϕ_1 e ϕ_2 , ocorreram na exclusão conjunta das observações e potencialmente influentes, e individualmente #61 e #69.

Tabela 9 – Parâmetros estimados por ML sem a presença individual e/ou conjunta de observações potencialmente influentes. Desvios padrões são apresentados entre parênteses.

Dados	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\phi}_1$	$\hat{\phi}_2$	IDE(%)
Todos	3,402 (0,362)	0,003 (0,006)	-0,388 (0,581)	-0,009 (0,010)	0,069 (0,099)	0,232 (0,053)	0,209 (0,159)	17,412 — — —
#(52, 61, 69)	3,054 (0,417)	0,009 (0,006)	0,370 (0,590)	-0,018 (0,012)	0,214 (0,101)	0,245 (0,047)	0,094 (0,110)	10,199 — — —
#52	2,969 (0,385)	-0,002 (0,007)	-0,557 (0,587)	0,003 (0,010)	0,135 (0,103)	0,232 (0,054)	0,208 (0,159)	17,364 — — —
#61	3,104 (0,390)	-0,002 (0,006)	-0,515 (0,585)	-0,003 (0,012)	0,169 (0,102)	0,239 (0,052)	0,173 (0,146)	15,452 — — —
#69	2,990 (0,360)	-0,003 (0,006)	-0,604 (0,579)	0,000 (0,010)	0,208 (0,098)	0,240 (0,047)	0,116 (0,118)	11,970 — — —

Fonte: Elaborada pelo autor (2022).

Com relação as estimativas dos parâmetros que definem a média $(\beta_i)'s, i = 0, 1, \dots, 4$, ocorreu variações para todos os casos, notadamente mais expressivas nas exclusões individuais, sobretudo as estimativa de β_1 em que sua influência na resposta muda de direção (vista pela alternância de sinal). O mesmo ocorre para β_3 que sua influência na resposta muda de direção na exclusão do ponto #52 e se anula para observação #69. Ainda, verifica-se alterações significativas para os erros padrões em todos dos casos. Além disso, fica evidente que a exclusão das observações potencialmente influentes reduziu, significativamente, a dependência espacial entre as amostras vista pela redução do *IDE*. Neste caso, destaca-se o caso da exclusão de todas as observações potencialmente influentes, e a retirada individual do ponto #69, em que o *IDE* reduziu-se do intervalo de classificação moderada para fraca dependência espacial. Portanto, a modelagem deve ser realizada com cautela sob a presença ou não da observação #69, pois a mesma também foi destacada como *outlier* no gráfico BoxPlot (Figura 9(b)).

2.7.2.2 Influência das observações potencialmente influentes na confecção dos Mapas de predição

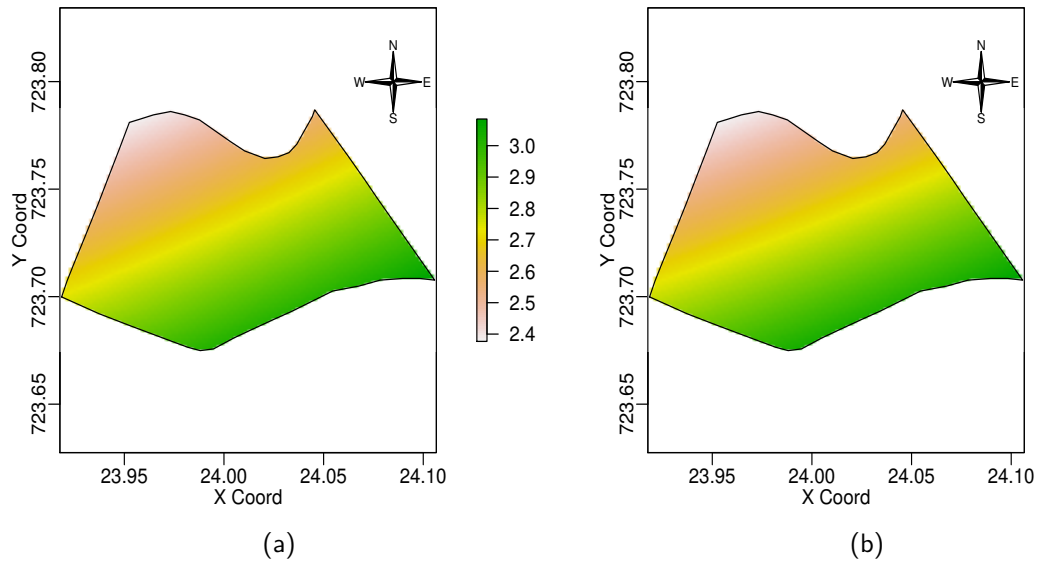
A partir das análise precedentes, e definição do modelo teórico de semivariograma, então, com o objetivo de reconstruir a área contínua em estudo, permitindo assim inferir sobre a variabilidade espacial do rendimento do grão de soja, realiza-se aqui, o processo de interpolação por Krigagem Universal (ver Diggle e Ribeiro (2007)), haja vista a inclusão de covariáveis,

conduzindo a construção dos mapas de variabilidade espacial (Figuras 11, 12 e 13).

A produtividade média da soja, quando considerada todas as observações (Figura 12(a)), obteve um valor máximo de 3,26 e um valor mínimo igual a 2,27 $t\ ha^{-1}$. Quando são excluídas as observações potencialmente influentes os valores máximo e mínimo foram 3,10 e 2,23, 3,16 e 2,21, 3,11 e 2,11, 3,35 e 2,45, nesta ordem, quando excluídas as observações #52, #61 e #69, observação #52, observação #61 e, observação #69, respectivamente. Nota-se, ainda, que os maiores rendimentos médios distribuem-se em parte da região sudoeste passando por sul com direção ao nordeste do mapa. Os valores menores concentraram-se, exclusivamente, na região norte com tendência para o noroeste. O menor valor mínimo pra produtividade foi quando excluída a observação #61 (Figura 13(b)), enquanto o maior valor máximo quando excluída a observação #69 (Figura 13). Com isso, verifica-se que em todos os casos, as observações potencialmente influentes, alteraram de forma significativa a geração dos mapas. Ademais, destaca-se a observação #69, que com sua exclusão foi a que apresentou maior diferença nos mapas com relação ao mapa com todas observações, visto pela região noroeste com alta concentração de valores pequenos.

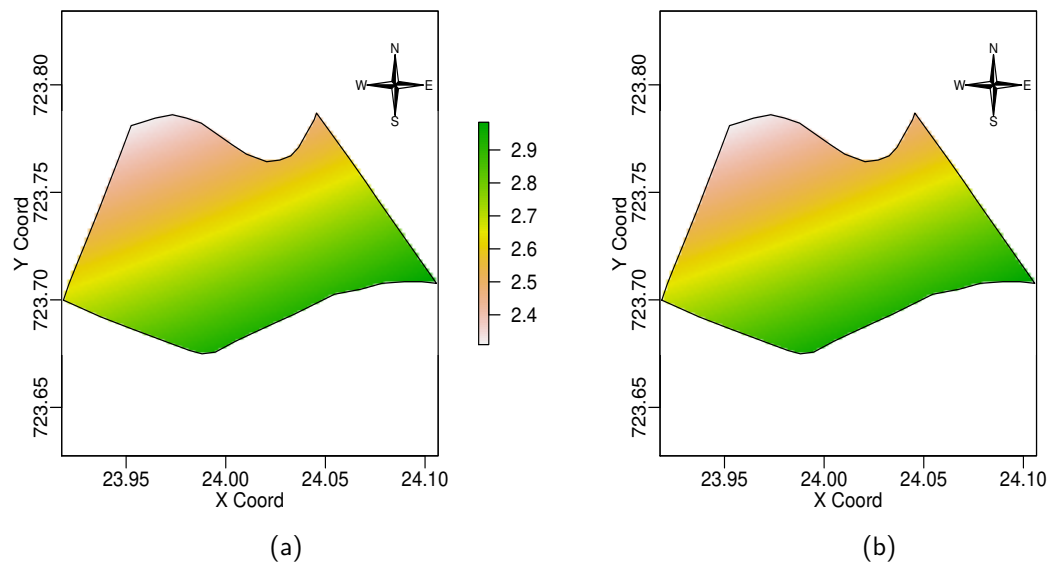
Dessarte, verifica-se que as observações consideradas como influentes exercem um papel importante na confecção dos mapas, visto que, a presença ou não destas, provocou alterações consideráveis nos mapas de previsão. Diante disto, a análise geoestatística da região em estudo deve ser realizada com cautela na presença ou não destas observações, sobretudo, nas regiões em que as mesmas foram amostradas. Portanto, a partir da análise dos mapas é possível localizar as zonas em que ocorrem as maiores e menores estimativas para o rendimento médio da soja, assim, poderão ser utilizados para gerenciar a variabilidade espacial do rendimento médio de soja da área em estudo, influenciando nas práticas de manejo.

Figura 11 – Mapas para os valores previstos de produtividade da soja em ($t\ ha^{-1}$) na safra 2016/2017 para a função de covariância de Wendland com suavização $\kappa = 3$, 12(a) com todos os pontos e 12(b) sem as observações #52, #61 e #69.



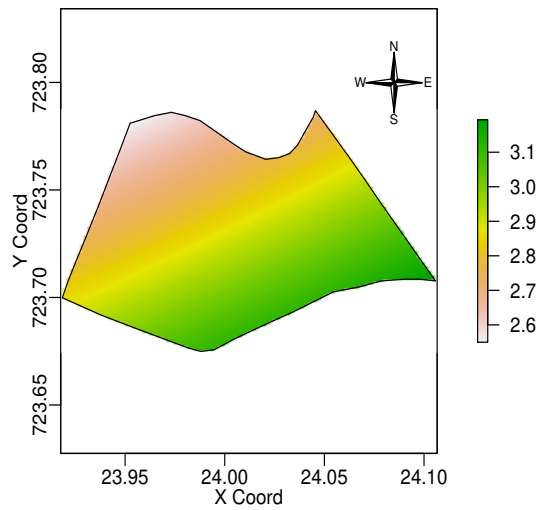
Fonte: Elaborada pelo autor (2022).

Figura 12 – Mapas para os valores previstos de produtividade da soja em ($t\ ha^{-1}$) na safra 2016/2017 para a função de covariância de Wendland com suavização $\kappa = 3$, 13(a) sem a observação #52, e 13(b) sem a observação #61.



Fonte: Elaborada pelo autor (2022).

Figura 13 – Mapas para os valores previstos de produtividade da soja em ($t \ ha^{-1}$) na safra 2016/2017 para a função de covariância de Wendland com suavização $\kappa = 3$ sem a observação #69.



Fonte: Elaborada pelo autor (2022).

2.8 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Em resumo, este Capítulo apresentou um estudo das funções de covariância pertencentes a família Wendland em uma aplicação geoestatística. Esta abordagem mostrou-se interessante, no sentido de que é importante estudar o comportamento de funções de covariância além das mais populares. Sua flexibilidade por meio de um parâmetro de suavização a colocou como uma forte concorrente da família Matérn. A aplicação por meio de conjunto de dados da produtividade da soja, mostrou que a metodologia aplicada foi eficiente para detectar pontos potencialmente influentes. Neste caso, os resultados mostraram que os pontos classificados como influentes afetaram as estimativas dos parâmetros que definem a estrutura de dependência espacial e influenciaram na construção dos mapas, isto é, a exclusão das observações ocasionaram alterações na distribuição espacial do rendimento médio da soja.

3 UM NOVO CRITÉRIO PARA SELEÇÃO DA MATRIZ DE CORRELAÇÃO DE TRABALHO EM MODELOS ESPACIAIS DE QUASI - VEROSSIMILHANÇA

3.1 RESUMO

Assim como na proposta original (para dados longitudinais), a modelagem das Equações de Estimação Generalizadas - EEG, aplicada em análise de dados espacialmente correlacionados, requer que seja especificada uma matriz de correlação. Geralmente nomeada matriz de correlação de trabalho, tem como característica influenciar as estimativas dos parâmetros de regressão, que são altamente sensíveis a escolha da matriz de correlação. Com isso, sua correta especificação pode garantir estimadores eficientes para os parâmetros de regressão. Diante disso, neste estudo propõe-se um novo método de seleção da matriz de correlação de trabalho, baseado no condicionamento da matriz de variância-covariância *naive*. A avaliação da performance do método é realizada por um estudo de simulações, utilizando as distribuições marginais da Normal, Poisson e Gama para dados espacialmente correlacionados. A especificação da estrutura de correlação é baseada em modelos de semivariogramas, utilizando as famílias Wendland, Matérn e modelo Esférico. Os resultados revelam que o critério proposto resultou em um melhor desempenho em comparação com os critérios concorrentes *quasi-likelihood under the independence model criterion* (QIC), *Crítério de Informação de Correlação* (CIC), e *Rotnizky and Jewell's criterion* (RJC), com relação as taxas de acerto da verdadeira estrutura de correlação espacial dos dados simulados. Um exemplo numérico de seleção da estrutura de correlação espacial apropriada foi mostrado utilizando os dados de precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil. Uma análise de diagnóstico é desenvolvida, em que são obtidas medidas de influência local sob os esquemas de perturbações no preditor linear e na matriz de correlação de trabalho. Para estas medidas considera-se as distribuições Binomial, Poisson, Binomial negativa, Gama e Normal inversa.

3.2 INTRODUÇÃO

O modelo de regressão normal foi, e ainda é, considerado uma ferramenta poderosa para modelagem e análise de dados. No entanto, seu emprego requer que sejam satisfeitas algumas suposições, como normalidade, independência e homoscedasticidade dos erros. Segundo Paula (2013) esta metodologia foi largamente utilizada para descrever diversos fenômenos

aleatórios, inclusive em situações que as suposições de normalidade não eram satisfeitas ou aplicáveis. E nestes casos, algum tipo de transformação era sugerida para que assim as suposições pré-estabelecidas fossem atendidas, apesar de à época existiam aplicações com modelos não lineares ou não normais.

Foi então que no início da década de 70 uma nova classe de modelos de regressão foi proposta por Nelder e Wedderburn (1972). Denominada de Modelos Lineares Generalizados (*Generalized Linear Models - Modelo linear generalizado (GLM)*), que de acordo com os autores incluem, como casos especiais, modelos de regressão linear e análise de variância e modelos de resposta multinomial para dados de contagem e alguns modelos em geral usados para dados de sobrevivência.

À vista disso, os GLM's consistem em expandir o leque de opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, além de permitir maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear η , (PAULA, 2013). Isto é, permitem dilatar as suposições inicialmente admitidas e estudar não somente as relações entre as variáveis explicativa e a variável resposta. Posto isto, fica evidente que a classe dos GLM's são capazes de modelar uma variedade de dados. Apesar disso, devido a sua limitação de exigir que os erros sejam independentes, os mesmos não são capazes de modelar dados correlacionados, incluindo dados com estrutura espacial, longitudinais ou multiníveis. Para contornar esta problemática, uma alternativa é utilizar os Modelos Lineares Generalizados Mistos ou modelos de quasi-verossimilhança.

Os Modelos Lineares Generalizados Mistos (Modelos lineares generalizados mistos (MGLM)) são extensões dos GLM's obtidos a partir da inclusão de variáveis aleatórias não observáveis, ou seja, os efeitos aleatórios no preditor linear, os quais são amplamente utilizados em análise de dados geoespaciais. Por outro lado, os modelos de quasi-verossimilhança, o qual será utilizado neste Capítulo, também podem ser estendidos para análise de dados geoespaciais, contudo, sem a especificação de uma distribuição. Para este caso, bastando indicar uma relação entre a média e a variância, utilizando alguma função de variância. Esta classe de modelos foi proposta por Wedderburn (1974) como uma generalização das ideias de Nelder e Wedderburn (1972) e, posteriormente estendidos por McCullagh (1983), Nelder e Pregibon (1987). Ademais, no que tange à análise de dados correlacionados os primeiros trabalhos remota de Stiratelli, Laird e Ware (1984) em que consideram um modelo linear logístico misto para análise de dados longitudinais. Os autores Zeger, Liang e Self (1985) propuseram através de extensões da regressão logística modelos que geram estimativas consistentes dos parâmetros de regressão e da

variância sob dependência no tempo e, Liang e Zeger (1986) desenvolveram uma técnica que permitem o estudo de variáveis aleatórias correlacionadas não gaussianas, nomeadas Equações de Estimação Generalizadas (EEG's), que podem ser utilizadas para estimação dos parâmetros nos modelos de quasi-verossimilhança. A partir destas ideias, Albert e McShane (1995) e Lin, Clayton et al. (2005) utilizaram as EEG's aplicadas a dados binários com dependência espacial, modelando uma matriz de correlação de trabalho usando uma função baseada na estrutura do semivariograma.

Portanto, as EEG's permitem modelar de forma explícita a estrutura da correlação dos dados, sendo assim possível ajustar a variabilidade entre as observações além de fornecer estimativas consistentes dos parâmetros de regressão. Desta forma, desde o advento dos modelos de quasi-verossimilhança e das EEG's, diversos trabalhos tem sido proposto na literatura. Por exemplo: Nelder e Pregibon (1987) estenderam a função de quasi-verossimilhança para permitir comparar as funções de variâncias. Huzii (1988) examinaram o comportamento da função condicional de quase - verossimilhança em séries temporais. Zeger, Liang e Albert (1988) utilizaram uma abordagem das EEG's para modelar dados longitudinais nos casos contínuos e discreto. Vos (1992) mostrou que as estimativas de máxima quasi - verossimilhança constituem um caso especial de uma família de estimadores do tipo f -divergência mínima. Chang (1999) mostrou a consistência para as estimativas de máxima quase - verossimilhança. Chang (2000) aplicam a teoria das EEG's para análise dos resíduos no estudo de dados longitudinais. Cantoni, Flemming e Ronchetti (2005) estabeleceu critérios de seleção de variáveis em modelos longitudinais marginais. Godambe e Heyde (2010) estabelecem uma definição geral de equações de estimativas de quasi - verossimilhança. Guan, Jalilian e Waagepetersen (2015) mostraram que os métodos de quase - verossimilhança são estatisticamente eficientes para dados de pontos espaciais. Nava et al. (2017) consideraram um modelo de distribuição marginal binomial na realização de um estudo de influência local com uma aplicação para dados com resposta espacialmente correlacionadas. Deng et al. (2017) propuseram um novo procedimento de quase-verossimilhança de segunda ordem para estimar os parâmetros em processos de pontos espaciais, no qual são obtidas funções de estimativa de primeira e segunda ordem combinadas linearmente com funções de peso apropriadas. Chu (2017) propõe o uso de estimativas de máxima quasi-verossimilhança para estimar modelos de correção de erros de grandes painéis de dados com covariáveis estacionárias, não estacionárias e erros espacialmente dependentes. Cerveckí et al. (2019) desenvolveram uma abordagem alternativa de quasi-verossimilhança para modelos funcionais. Sato e Matsuda (2020) empregam uma extensão dos modelos de

heteroscedasticidade condicional autorregressiva generalizada (GARCH), utilizando como procedimento de estimação dos parâmetros funções de quasi-verossimilhança. Sutradhar e Oyet (2020) utilizam funções de quasi-verossimilhança generalizada para inferência em modelo de correlação de dados binários espaciais usando uma abordagem de modelo *logistic* misto, em que correlações de pares são calculadas acomodando dentro e entre correlações para respostas em pares. Manuel e Scalon (2020) aplicam as EEG's no estudo de dados de rede espacial modelando a matriz de correlação de trabalho espacial usando o índice de Moran e a matriz de peso espacial. Li, Cai e Li (2021) apresentam os modelos de quasi-verossimilhança para estimativa e inferência de parâmetros desconhecidos em modelos lineares de efeitos mistos de alta dimensão.

Contudo, apesar de ser amplamente utilizada, a abordagem das EEG's ainda tem sido pouco explorada no contexto da análise de dados espacialmente correlacionados. Sua aplicação requer, *a priori*, que seja especificada uma matriz de correlação, comumente nomeada de correlação de trabalho. Esta matriz é geralmente desconhecida e representada por $\mathbf{R}(\varphi)$, em que φ , também desconhecido, é um vetor paramétrico de correlação que caracteriza completamente a matriz $\mathbf{R}$. De modo geral, as EEG's tem como ponto forte gerar estimativas consistentes para os parâmetros de regressão, mesmo que a matriz de correlação de trabalho não seja especificada corretamente, ver Venezuela, Sandoval e Botter (2011). No entanto, uma especificação da matriz de correlação de trabalho que se aproxima da verdadeira matriz correlação que gerou os dados, poderá melhorar a eficiência dos estimadores dos parâmetros, sobretudo em pequenas amostras.

Nesta vertente, a literatura dispõe de vários critérios para auxiliar na escolha da melhor matriz de correlação de trabalho. Dentre os mais populares estão os *Rotnizky and Jewell's criterion* (RJC), *quasi-likelihood under the independence model criterion* (QIC), e *Crítério de Informação de Correlação* (CIC). O *Rotnizky and Jewell's criterion* foi inicialmente proposto por Rotnitzky e Jewell (1990), que sob a hipótese da verdadeira matriz de correlação, avaliaram o comportamento assintótico da estatística de teste Wald. Posteriormente, e embasados nestas ideias, Hin, Carey e Wang (2007) descreveram o critério RJC para avaliar seu desempenho em comparação com o QIC na seleção da estrutura da matriz de correlação de trabalho, utilizando simulações com resposta gaussiana ou binomial. Este último critério (QIC) foi apresentado por Pan (2001). Para seu desenvolvimento, o autor baseou-se no conhecido critério para seleção de modelos, Critério de Akaike (AIC), propondo uma modificação para acomodar a seleção de modelos sob EEG. Apesar de ser criado para seleção de modelos em EEG, tem sido amplamente

utilizado como critério de escolha da matriz de correlação de trabalho. E por isso, os autores Hin e Wang (2009) sugeriram utilizar metade do segundo termo do QIC, dando origem ao CIC. Justificando-o que o primeiro termo do QIC, por utilizar apenas a quasi-verossimilhança, não exerce influência sobre informações da matriz de correlação de trabalho.

Estes critérios, bem como alguns outros não tão populares, foram estabelecidos no contexto da análise de dados longitudinais. Ademais, pouco se conhece sobre o desempenho relativo, de cada critério, com relação a seleção de uma estrutura da matriz de correlação de trabalho ótima para melhorar a eficiência dos estimadores de regressão. Quando os dados são georreferenciados com uma estrutura de correlação gerada a partir de algum modelo de semivariograma, nada se sabe sobre o desempenho destes critérios.

De modo geral, os autores Zhao, Prentice e Self (1992) fornecem resultados importantes sobre a perda de eficiência das estimativas dos parâmetros de regressão sob EEG, quando a especificação da matriz de covariância da variável resposta é incorreta. Nesta mesma linha, Sutradhar e Das (1999) os autores examinaram a perda de eficiência dos estimadores de regressão devido à especificação incorreta da estrutura de correlação, baseados no trabalho original de Liang e Zeger (1986) com covariáveis em níveis de *cluster*. Posteriormente, Sutradhar e Das (2000) estudaram a problemática realizando uma avaliação computacional, com base em uma matriz de correlação de trabalho independente. Com isso, os autores compararam os estimadores sob independência e sob uma estrutura de correlação AR(1). Assim, concluíram que a eficiência dos estimadores está fortemente vinculada a correta especificação da matriz de correlação de trabalho.

No âmbito da análise de dados espacialmente correlacionados, a abordagem por EEG inicia-se definindo um vetor $\mathbf{Z} = (Z(\mathbf{s}_1) = z_1, \dots, Z(\mathbf{s}_n) = z_n)^\top$ de variáveis respostas observadas em diferentes localizações $\mathbf{s}_1, \dots, \mathbf{s}_n \in \mathbf{S} \subset \mathbb{R}^2$. Então, a partir da proposta de Liang e Zeger (1986) a matriz de covariância espacial é definida por $\mathbf{V} = \mathbf{A}^{1/2} \mathbf{R}(\boldsymbol{\varphi}) \mathbf{A}^{1/2}$, desde que se assuma um estimador consistente para os parâmetros de correlação espacial $\boldsymbol{\varphi}$, e $\mathbf{A} = \text{diag}(c(\sigma_1^2), \dots, c(\sigma_n^2))$ é uma matriz de variâncias dos dados com c definindo a dispersão. Aqui, considera-se $c = 1$. Desta forma, as estimativas de quasi-verossimilhança, para os parâmetros do modelo de regressão, são obtidas pela solução da função score

$$\mathbf{U} = \mathbf{D}^\top \mathbf{V}^{-1}(\mathbf{Z} - \boldsymbol{\mu}),$$

em que o vetor de médias tem elementos $\mu_i = g(\mathbf{x}_i^\top \boldsymbol{\beta})$, para alguma função $g(\cdot)$ monótona, diferenciável e invertível. E $\mathbf{x}$ é um vetor de ordem $p \times 1$ que representam as entradas das

linhas da matriz de covariáveis e, β de ordem $p \times 1$ é um vetor de parâmetros do modelo de regressão. D é uma matriz de derivadas da média com relação a β .

Devido a forma não fechada da função escore, sua solução é realizada por algum método iterativo. Então, dado um estimador consistente para os parâmetros da estrutura de correlação, um estimador consistente para os parâmetros de regressão, disponível na literatura como por exemplo nos trabalhos de Prentice (1988), Albert e McShane (1995), Venezuela, Botter e Sandoval (2007) e Nava et al. (2017), é dado por:

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + \left[\left(\hat{D}^{(m)} \right)^\top \left(\hat{V}^{(m)} \right)^{-1} \left(\hat{D}^{(m)} \right) \right]^{-1} \left(\hat{D}^{(m)} \right)^\top \left(\hat{V}^{(m)} \right)^{-1} \left(Z - \hat{\mu}^{(m)} \right). \quad (3.1)$$

Com isso, um estimador robusto (ou *sandwich*) e consistente para variância é:

$$\hat{V}_{rob} = \left(\hat{D}^\top \hat{V}^{-1} \hat{D} \right)^{-1} \left(\hat{D}^\top \hat{V}^{-1} Cov(Z) \hat{V}^{-1} \hat{D} \right) \left(\hat{D}^\top \hat{V}^{-1} \hat{D} \right)^{-1},$$

em que $Cov(Z) = (Z - \hat{\mu})(Z - \hat{\mu})^\top$. Note que, se a matriz de trabalho coincide ou é próxima da verdadeira matriz de correlação de trabalho, então, $\hat{V}$ aproxima-se ou coincide com $Cov(Z)$. Neste caso, o estimador de variância-covariâncias robusto reduz-se a um estimador ótimo, comumente nomeado de *naïve*, e dado por:

$$\hat{V}_N = \left(\hat{D}^\top \hat{V}^{-1} \hat{D} \right)^{-1}.$$

Portanto, qualquer que seja a matriz de correlação de trabalho, dado um vetor não-nulo x , temos que

$$x^\top (\hat{V}_{rob} - \hat{V}_N) x \geq 0,$$

isto é, a diferença entre os estimadores robusto e *naïve* resulta em uma matriz definida não-negativa. Logo, uma matriz de correlação de trabalho mal especificada poderá resultar em uma significativa perda de eficiência. Mas, este problema poderá ser controlado pelo condicionamento da matriz $\hat{V}_N$, haja vista sua influência no controle da propagação dos erros de arredondamento na Equação iterativa (3.1).

O condicionamento ou número de condição matricial está fortemente relacionado ao conceito de eficiência numérica, sobretudo em problemas que envolvem inversão matricial. Para este estudo, nos motivamos no fato de que o estimador robusto depende da inversão de termos equivalentes ao estimador $\hat{V}_N$. Neste caso, se $\hat{V}_N$ é mal condicionada, então, $\hat{V}_{rob}$ tende a ser mal condicionado. Ainda, com relação a eficiência numérica do estimador descrito na Equação (3.1), o processo iterativo depende da inversão matricial repetidas vezes até a convergência. Então, o mal condicionamento da matriz $\hat{V}_N$ poderá prejudicar a eficiência numérica

do estimador $\hat{\beta}$, podendo ainda ser dilatada pelos erros de arredondamento e aproximação numérica.

Com isso, neste estudo, propomos um novo critério para auxiliar na seleção da matriz de correlação de trabalho espacial em EEG, com base no número de condição da matriz de variâncias-covariâncias *naive*. Particularmente, esta proposta é direcionada a análise de dados espacialmente correlacionados. O critério é construído similarmente ao RJC, com a diferença que ao invés do traço das matrizes de variâncias-covariâncias estimada, introduzimos os conceitos de norma e condicionamento matricial. Além disso, adicionalmente, desenvolvemos algumas medidas de influência local para vários membros da família exponencial de distribuições, até então não disponíveis na literatura no que concerne a dados espacialmente correlacionados. Para avaliar o desempenho do critério proposto, realizamos um extenso estudo de simulação e comparamos os resultados com três critérios bastante conhecidos na literatura. A originalidade deste trabalho está no fato de que não estão disponíveis na literatura critérios que auxiliem na escolha da matriz de correlação de trabalho, ou seja, matriz de correlação geradas a partir de algum modelo de semivariograma, sobretudo com o conceito de condicionamento matricial.

3.3 MODELO DE QUASI - VEROSSIMILHANÇA CORRELACIONADO

Seja $\mathbf{Z} = (Z_1(\mathbf{s}_1), \dots, Z_n(\mathbf{s}_n))^T$ um vetor de variáveis respostas observadas nas localizações $\mathbf{s}_1, \dots, \mathbf{s}_n$ pertencentes ao espaço $\mathbb{R}^2$. Por simplicidade será usada a notação $Z(\mathbf{s}_i) = z_i$ $i = 1, \dots, n$ para indicar a variável resposta observada na localização $\mathbf{s}_i$. Então, através de uma abordagem por equações de estimação generalizadas, para modelar a estrutura marginal da média, assume-se que a distribuição marginal de z_i pertença a família exponencial, isto é:

$$f(z_i) = \exp \{c [z_i \theta_i - a(\theta_i)] + b(z_i; \theta_i)\},$$

em que $E(z_i) = \frac{\partial a(\theta_i)}{\partial \theta_i} = \mu_i$ e $Var(z_i) = c^{-1} A_i$, com $A_i = \frac{\partial^2 a(\theta_i)}{\partial \theta_i^2}$. Sem perda de generalidade, considera-se $c = 1$. Ainda, define-se $g(\cdot)$ como sendo uma função monótona, diferenciável e invertível, tal que:

$$g(\mu_i(\mathbf{s}_i)) = \eta_i = \mathbf{x}_i \boldsymbol{\beta},$$

em que $\mathbf{x}_i = (x_i, \dots, x_p)$, e $\boldsymbol{\beta}$, de ordem $p \times 1$ é um vetor de parâmetros desconhecidos, com $\mathbf{X}$ uma matriz de planejamento associada as covariáveis.

Desta forma, sob a suposição de independência, tem-se os Modelos Lineares Generalizados (MLG's) propostos por Nelder e Wedderburn (1972). Neste caso, a função score e matriz de

informação, respectivamente, são definidas por:

$$\mathbf{U}_\beta = \mathbf{D}^\top \mathbf{A}^{-1}(\mathbf{z} - \boldsymbol{\mu}), \quad (3.2)$$

$$\mathbf{K}_\beta = \mathbf{D}^\top \mathbf{A}^{-1} \mathbf{D}, \quad (3.3)$$

em que $\mathbf{D} = \text{diag} \left(\frac{\partial \mu_1}{\partial \eta_1}, \dots, \frac{\partial \mu_n}{\partial \eta_n} \right)$, e $\mathbf{A} = \text{diag} (A_1, \dots, A_n)$ é a matriz da função de variâncias. O leitor interessado por mais detalhes pode consultar os trabalhos de Artes (1997) e Venezuela, Botter e Sandoval (2007).

Dito isto, considerando que não se conhece a distribuição de z_i , mas que A_i é conhecida como função da média, então os autores Wedderburn (1974) definiram os modelos de Quasi-verossimilhança (QV). Esta classe de modelos que posteriormente foram explanadas por (MCCULLAGH; NELDER, 1989), ostentam como principal característica a construção de uma pseudo - verossimilhança, não exigindo de suposições acerca da classe de distribuição a qual os dados pertencem, além de gerar estimadores eficientes. Então, a ideia por trás desta metodologia é encontrar uma quantidade que carregue em si a mesma informação que uma função de verossimilhança usual, apenas com a especificação da relação entre a média e a variância das observações sob a suposição de uma estrutura de correlação. Neste caso, as Equações (3.2) e (3.3), desde que se ignore a correlação existente entre as observações, são nomeadas de função quasi-escore e quasi-informação, respectivamente. Sendo assim, para incorporar a correlação Liang e Zeger (1986) e Zeger e Liang (1986) definiram as Equações de Estimação Generalizadas (EEG), introduzindo uma matriz de correlação denominada matriz de trabalho $\mathbf{R}(\boldsymbol{\varphi})$, em que $\boldsymbol{\varphi}$ é um vetor de parâmetros de correlação que não depende de β . Desta forma, as EEG's é vista na literatura como uma extensão natural das equações das equações de quasi-verossimilhança, sendo os termos que definem a correlação tratados como uma perturbação. Neste cenário, os autores definem a matriz de variâncias-covariâncias como:

$$\mathbf{V} = \mathbf{A}^{1/2} \mathbf{R}(\boldsymbol{\varphi}) \mathbf{A}^{1/2}, \quad (3.4)$$

desde que $\mathbf{R}$ seja a verdadeira matriz de correlação de $\mathbf{z}$. Portanto, as Equações 3.5 e 3.6 são reescritas como:

$$\mathbf{U}_\beta^* = \mathbf{D}^\top \mathbf{V}^{-1}(\mathbf{z} - \boldsymbol{\mu}), \quad (3.5)$$

$$\mathbf{K}_\beta^* = \mathbf{D}^\top \mathbf{V}^{-1} \mathbf{D}. \quad (3.6)$$

Note que, quando $\mathbf{R}$ é igual a matriz identidade as Equações (3.5) e (3.6) recuperam as Equações (3.2) e (3.3).

A EEG definida a partir da função quasi-escore (3.5), depende da matriz de variâncias-covariâncias $\mathbf{V}$, a qual depende do vetor de parâmetros φ associado a matriz de correlação espacial $\mathbf{R}$. Diante disso, os autores Liang e Zeger (1986) utilizam o método dos momentos para escrever os parâmetros de correlação e escala em função dos resíduos de Pearson. Então, aplica-se algum método iterativo na solução da quasi-verossimilhança para se estimar β e um método robusto para estimar φ . Um estimador consistente e assintoticamente normal, obtido por uma modificação do algoritmo escore de Fisher, dado um estimador consistente de φ , está disponível na literatura e, por exemplo pode ser encontrado nos trabalhos de Prentice (1988), Venezuela, Botter e Sandoval (2007), Nava et al. (2017), o qual é dado pela Equação (3.7).

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + \left[(\hat{\mathbf{D}}^{(m)})^\top (\hat{\mathbf{V}}^{(m)})^{-1} (\hat{\mathbf{D}}^{(m)}) \right]^{-1} (\hat{\mathbf{D}}^{(m)})^\top (\hat{\mathbf{V}}^{(m)})^{-1} (\mathbf{Z} - \hat{\mu}^{(m)}), \quad (3.7)$$

em que $m = 0, 1, 2, \dots$ é o passo de iteração. Desta forma, em cada passo, estimativas de φ devem ser fornecidas e aplicadas para atualizações das estimativas de $\mathbf{D}$, $\mathbf{V}$ e μ .

No artigo original de Liang e Zeger (1986), os autores advogam que se a matriz de correlação de trabalho é especificada corretamente, então, um estimador para matriz de variâncias-covariâncias, associada as estimativas dos parâmetros de regressão é consistente, sendo obtido por:

$$\hat{\mathbf{V}}_N(\hat{\beta}) = (\hat{\mathbf{D}}^\top \hat{\mathbf{V}}^{-1} \hat{\mathbf{D}})^{-1}. \quad (3.8)$$

Contudo, na prática a verdadeira matriz de correlação é em geral desconhecida, portanto, deve ser estimada. Neste caso, um estimador consistente para matriz de variâncias-covariâncias é dado por:

$$\hat{\mathbf{V}}_{rob}(\hat{\beta}) = (\hat{\mathbf{D}}^\top \hat{\mathbf{V}}^{-1} \hat{\mathbf{D}})^{-1} (\hat{\mathbf{D}}^\top \hat{\mathbf{V}}^{-1} Cov(\mathbf{z}) \hat{\mathbf{V}}^{-1} \hat{\mathbf{D}}) (\hat{\mathbf{D}}^\top \hat{\mathbf{V}}^{-1} \hat{\mathbf{D}})^{-1}, \quad (3.9)$$

em que $Cov(\mathbf{z}) = (\mathbf{z} - \hat{\mu})(\mathbf{z} - \hat{\mu})^\top$. Os estimadores (3.8) e (3.9) são nomeados de *naive* e *robusto*, respectivamente.

3.3.1 Estimação dos parâmetros de correlação espacial

Um estimador para o vetor de parâmetros associado a correlação de trabalho depende de sua estrutura, a qual pode ser especificada de diferentes formas, dentre elas: autoregressiva de primeira ordem, uniforme, não estruturada, M-dependente, espacial, dentre outras. Na hipótese

de o pesquisador possuir alguma informação sobre a estrutura de correlação dos dados, uma alternativa é utilizar uma matriz de correlação do tipo fixa. O que é a proposta aqui. Portanto, utilizamos um estimador consistente sob um modelo de média e semivariograma especificado corretamente estabelecido por Albert e McShane (1995), a partir de um processo iterativo (ver Equação (3.11)).

Destarte, seja $\varphi = (\phi_1, \phi_2, \phi_3)^\top$ com ϕ_1, ϕ_2 e ϕ_3 denotando o efeito pepita, contribuição ou variância do processo e parâmetro de escala, respectivamente. O parâmetro ϕ_1 é visto como variações em microescala, isto é, possíveis erros de medições. Por outro lado, ϕ_2 mensura a variância do processo, enquanto ϕ_3 controla o decaimento da curva de correlação além de definir a distância para os quais as observações estão espacialmente correlacionadas. Logo, considerando a matriz de variâncias-covariâncias definida pela Equação (3.4), então, a matriz de correlação espacial $\mathbf{R}$, de ordem $n \times n$, aqui é definida por:

$$\mathbf{R}(\varphi) = \begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \vdots & & \ddots & \vdots \\ r_{n1} & \dots & r_{n(n-1)} & 1 \end{pmatrix}.$$

Com isso, considerando h_{ij} como sendo a distância euclidiana entre as localizações espaciais $\mathbf{s}_i$ e $\mathbf{s}_j$, $i, j = 1, \dots, n$, o (i, j) -ésimo elemento da matriz $\mathbf{R}$, com $i \neq j$ é dado por:

$$r_{ij} = 1 - \phi_1 + \phi_2 \left[1 - \rho_{\mu, \kappa} \left(\frac{h_{ij}}{\phi_3} \right) \right],$$

em que $\rho : \mathbb{R}^2 \rightarrow \mathbb{R}$ define alguma função de correlação espacial. Por exemplo, para família Wendland generalizada com parâmetro de suavização κ , temos:

$$\rho_{\mu, \kappa}(h) = \begin{cases} \frac{1}{B(2\kappa, \mu+1)} \int_h^1 x (x^2 - h^2)^{\kappa-1} (1-x)^\mu dx, & \text{se } 0 \leq h < 1 \\ 0, & \text{se } h \geq 1, \end{cases}$$

em que $B(\cdot)$ denota a função Beta. Considera-se que $\kappa \in \mathbb{Z}_+$, o qual é denominado parâmetro de suavização. O parâmetro μ caracteriza a concavidade da curva e pela relação (3.10) define a positividade da matriz correlação gerada.

$$\mu \geq \frac{d+1}{2} + \kappa, \quad \text{com } d = 2. \quad (3.10)$$

A função (3.3.1) e a relação (3.10) foram definidas no Capítulo 2, com maiores detalhes nos trabalhos de Gneiting (2002a), Zastavnyi (2006) e Bevilacqua et al. (2019). Os dois

primeiros autores mostram que ρ é uma função positiva definida, em um espaço d -dimensional ($d = 1, 2, 3$) se, e somente se a condição (3.10) é satisfeita.

Com isso, as estimativas dos parâmetros que definem a estrutura de correlação podem ser obtidos pelo método dos mínimos quadrados generalizados ou pelo estimador consistente definido por Albert e McShane (1995). Para o segundo caso, as estimativas são obtidas por um processo iterativo dado por (3.11), dado o estimador consistente de β .

$$\hat{\varphi}^{(m+1)} = \hat{\varphi}^{(m)} + \left[\left(\frac{\partial \mathbf{v}}{\partial \hat{\varphi}} \right)^\top \hat{\mathbf{W}}^{-1} \frac{\partial \mathbf{v}}{\partial \hat{\varphi}} \right]^{-1} \left(\frac{\partial \mathbf{v}}{\partial \hat{\varphi}} \right)^\top \hat{\mathbf{W}}^{-1} (\hat{\boldsymbol{\sigma}} - \hat{\mathbf{v}}), \quad (3.11)$$

em que $\boldsymbol{\sigma}$ e $\mathbf{v}$ são os $\frac{n(n-1)}{2}$ elementos abaixo da diagonal principal das matrizes de covariância da amostral $Cov(\mathbf{z}) = (\mathbf{z} - \hat{\boldsymbol{\mu}})(\mathbf{z} - \hat{\boldsymbol{\mu}})^\top$ e $\mathbf{V}$, respectivamente. Enquanto $\mathbf{W}$ é uma matriz diagonal que define a variância de trabalho de $\boldsymbol{\sigma}$, com elementos, sugeridos por Prentice (1988), dados por

$$Var(\sigma(i, j)) = (1 - 2\boldsymbol{\mu}(\mathbf{s}_i))(1 - 2\boldsymbol{\mu}(\mathbf{s}_j))v_{ij} + v_{ii}v_{jj} - v_{ij}^2,$$

sendo v_{ij} o (i, j) -ésimo elemento da matriz $\mathbf{V}$.

3.3.2 Critérios para seleção da matriz de correlação

O custo da flexibilidade da existência de várias matrizes de trabalho concentra-se na perda de eficiência dos estimadores $\hat{\beta}$ de β e $\hat{\mathbf{V}}_{rob}$, ao passo que se especifique uma matriz de correlação distante da verdadeira. Portanto, é necessário considerar alguma medida que auxilie na escolha da melhor matriz de correlação a ser especificada. Diante disso, são vários critérios para seleção da melhor matriz de correlação no contexto das EEG disponíveis na literatura. Dentre os mais populares estão o *quasi-likelihood under the independence model criterion* (QIC), *Critério de Informação de Correlação* (CIC), e *Rotnizky and Jewell's criterion* (RJC), propostos inicialmente por Pan (2001), Hin e Wang (2009) e Rotnitzky e Jewell (1990), Hin, Carey e Wang (2007), respectivamente.

A ideia do QIC é realizar uma modificação adequada no popular critério de informação de Akaike (AIC). Sua medida é obtida pela equação (3.12).

$$QIC = -2Q(\hat{\boldsymbol{\mu}}) + 2tr(\hat{\mathbf{V}}_I^{-1}\hat{\mathbf{V}}_{rob}), \quad (3.12)$$

em que $Q(\cdot)$ denota a função de quasi-verossimilhança, $\hat{\mathbf{V}}_I$ a matriz de variâncias-covariâncias para os estimadores dos parâmetros sob independência e, $\hat{\mathbf{V}}_{rob}$ é a matriz de variâncias-covariância estimada a partir do estimador robusto (3.9).

O critério de informação da matriz de correlação (CIC) é obtido a partir da metade do segundo termo do QIC, ver Equação (3.13).

$$CIC = tr \left(\hat{\mathbf{V}}_I^{-1} \hat{\mathbf{V}}_{rob} \right). \quad (3.13)$$

O critério RJC, em que originalmente estabelecido por Rotnitzky e Jewell (1990) é uma extensão do teste qui-quadrado que examina-se o comportamento assintótico da estatística de Wald generalizada. Contudo, a partir de uma especificação correta da matriz de correlação $\mathbf{H}_n$, Carey e Wang (2007) generalizou as ideias dos autores acima citados, e estabeleceram o critério RJC por 3.14.

$$\begin{aligned} RJ_1 &= tr \left(\hat{\mathbf{V}}_{rob} \hat{\mathbf{V}}_{Naive}^{-1} \right) / p, \\ RJ_2 &= tr \left(\hat{\mathbf{V}}_{rob} \hat{\mathbf{V}}_{Naive}^{-1} \right)^2 / p, \\ RJC &= \sqrt{(RJ_1 - 1)^2 + (RJ_2 - 1)^2} \end{aligned} \quad (3.14)$$

em que $\hat{\mathbf{V}}_{Naive}$ é a matriz de variâncias-covariâncias estimada a partir do estimador consistente (3.8), e p é o número de parâmetros do modelo de regressão.

Além destes, existem outros critérios disponíveis na literatura, como por exemplo: Shults-Chaganty's (SC) proposto por Shults e Chaganty (1998) no qual baseia-se nos mínimos quadrados ponderados utilizando $\mathbf{V}^{-1}$ como matriz de pesos; a estatística de soma de quadrados residual predita (PRESS), proposta por Inan, Latif e Preisser (2018) obtida como uma modificação na matriz de peso no critério SC; dentre outros. Todas estas medidas podem ser utilizadas como critério de seleção da matriz de correlação de trabalho.

3.4 CONDITION-NUMBER-BASED CRITERION: CRITÉRIO PARA SELEÇÃO DA MATRIZ DE CORRELAÇÃO DE TRABALHO

Como é bem discutido na literatura, o método EEG proporciona ótimos resultados para estimativas do modelo regressão. No entanto, o desafio na aplicação do método está em uma boa especificação para matriz de correlação de trabalho. No caso de análise de dados espaciais a matriz de correlação espacial. A escolha de uma matriz de correlação espacial adequada é de extrema importância, pois uma matriz inadequada poderá gerar estimativas

pouco confiáveis haja vista o aumento do viés de estimação. Portanto, é necessário que se estabeleça critérios que auxiliem na escolha da matriz de correlação espacial, assim conduzir à estimativas consistentes e, consequentemente evitar a perda de eficiência na estimação dos parâmetros de regressão. Como mencionado na Seção 3.3, dentre os critérios mais populares estão o *QIC*, *CIC* e *RJC*. No entanto, estes critérios levam em consideração apenas os elementos da diagonal principal do produto das matrizes de variâncias-covariâncias, obtidas pelo estimador robusto e sob independência ou robusto e *naive*. Ademais, de acordo com Hin, Carey e Wang (2007) quando se utiliza o critério *RJC*, sob a especificação de uma correta matriz de correlação, o produto $\hat{\mathbf{V}}_{Naive}^{-1} \times \hat{\mathbf{V}}_{rob}$ se aproxima de uma matriz identidade. Contudo, os elementos fora da diagonal se aproximam de zero, mas não necessariamente são iguais a zero. Da mesma forma, os elementos da diagonal não são necessariamente iguais a um. Diante disso, sente-se a necessidade de um critério que carregue toda a informação presente nas matrizes variâncias-covariâncias bem como da matriz de covariância espacial.

São vários os critérios de otimalidade disponíveis, que tem como objetivo fornecer uma medida de ajuste para auxiliar na seleção de modelos de regressão. Em geral, tais critérios são classificados como os que se baseiam em informações (particularmente da matriz de variância-covariâncias) e, os que se baseiam em distâncias. Para o primeiro caso, os mais populares são os D-otimalidade e A-otimalidade, que baseiam-se na variância generalizada (determinante) e na variância total (traço) da matriz de variâncias-covariâncias dos parâmetros de regressão, respectivamente. No segundo caso, a medida é construída a partir de uma cobertura de pontos (nuvem de espalhamento) em um espaço p -dimensional, em que p é o número de termos do modelo. Contudo, a necessidade de armazenamento de uma matriz de distâncias torna este critério muito custoso computacionalmente, o tornando pouco utilizado na prática. Para uma leitura com mais detalhes sobre estes critérios, o leitor pode consultar Wong (1994) e Atkinson, Donev e Tobias (2007).

Assim sendo, nota-se que os critérios para seleção da matriz correlação de trabalho dependem do estimador de variâncias robusto, que por sua vez depende do processo de inversão matricial. Portanto, é importante considerar como os erros de arredondamento se propagam, ou seja, avaliar como pequenas variações nos elementos de $\mathbf{V}$ e $\mathbf{V}^{-1}$ influenciam nos resultados da estatística de seleção. Diante disso, a proposta aqui, é estabelecer um critério baseado em uma modificação nas expressões do critério *RJC*, introduzindo a norma e número de condição das matrizes $\hat{\mathbf{V}}_{rob}$ e $\hat{\mathbf{V}}_{Naive}$, respectivamente. Desta maneira, espera-se capturar toda informação trazida da matriz de correlação espacial para estimativas das matrizes de

variâncias-covariâncias.

A introdução do número de condição de uma matriz tem como objetivo avaliar a sensibilidade da matriz de variâncias-covariâncias. O mal condicionamento de uma matriz está relacionado a quase singularidade da matriz e, sua definição é facilmente encontrada nos principais textos sobre análise numérica como sendo:

$$Cond(\mathbf{B}) = \|\mathbf{B}\| \|\mathbf{B}^{-1}\|,$$

em que $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$ define uma norma matricial induzida por uma norma vetorial, aqui a norma euclidiana. Portanto, a grandeza do número de condição da matriz $\mathbf{B}$ revela a sensibilidade da solução do problema as perturbações na matriz de correlação de trabalho. O leitor interessado poderá consultar maiores detalhes em Belsley, Kuh e Welsch (2005) e Chapra e Canale (2016). O Primeiro autor utiliza-se o número de condição para auxiliar no diagnóstico de colinearidade entre as variáveis de regressão, enquanto que o segundo à utiliza no contexto geral da análise numérica.

De mais a mais, destaca-se que o estimador de variâncias-covariâncias robusto assume propriedades assintóticas, portanto, não é válido sempre. Contudo, se o modelo marginal de regressão e a matriz de correlação de trabalho estão corretamente especificados, então, estimador robusto reduz-se ao estimador *naive*. Com isso, aplicando a norma em ambos os lados da equação iterativa ((3.7)) para estimativas dos parâmetros de regressão, temos que

$$\begin{aligned} \|\hat{\beta}^{(m+1)} - \hat{\beta}^{(m)}\| &= \left\| \left[(\hat{\mathbf{D}}^{(m)})^\top (\hat{\mathbf{V}}^{(m)})^{-1} (\hat{\mathbf{D}}^{(m)}) \right]^{-1} (\hat{\mathbf{D}}^{(m)})^\top (\hat{\mathbf{V}}^{(m)})^{-1} (\mathbf{Z} - \hat{\mu}^{(m)}) \right\| \\ &\leq \|\hat{\mathbf{V}}_{Naive}\| \left\| (\hat{\mathbf{D}}^{(m)})^\top \right\| \left\| (\hat{\mathbf{V}}^{(m)})^{-1} \right\| \|\mathbf{Z} - \hat{\mu}^{(m)}\|. \end{aligned}$$

Então, como $\hat{\beta}^{(m+1)}$ é solução do sistema de EEG, podemos afirmar que

$$\hat{\beta}^{(m+1)} > (\hat{\mathbf{D}}^{(m)})^\top (\hat{\mathbf{V}}^{(m)})^{-1} (\mathbf{Z} - \hat{\mu}^{(m)}).$$

Portanto, utilizando uma manipulação algébrica temos que

$$\begin{aligned} \frac{\|\hat{\beta}^{(m+1)} - \hat{\beta}^{(m)}\|}{\|\hat{\beta}^{(m+1)}\|} &\leq \frac{\|\hat{\mathbf{V}}_{Naive}\| \|\hat{\mathbf{V}}_{Naive}^{-1}\| \left\| (\hat{\mathbf{D}}^{(m)})^\top \right\| \left\| (\hat{\mathbf{V}}^{(m)})^{-1} \right\| \|\mathbf{Z} - \hat{\mu}^{(m)}\|}{\|\hat{\mathbf{V}}_{Naive}^{-1}\| \left\| (\hat{\mathbf{D}}^{(m)})^\top \right\| \left\| (\hat{\mathbf{V}}^{(m)})^{-1} \right\| \|\mathbf{Z} - \hat{\mu}^{(m)}\|} \\ &= Cond(\hat{\mathbf{V}}_{Naive}) \frac{1}{\|\hat{\mathbf{V}}_{Naive}^{-1}\|}. \end{aligned}$$

Desta forma, temos um majorante para o erro relativo das estimativas dos parâmetros do modelo marginal de regressão sob EEG. Portanto, o erro relativo é proporcional ao inverso da

norma da matriz estimada para variâncias-covariâncias, sob o estimador *naive*. Neste caso, o número de condição é um fator de proporcionalidade, isto é, quantifica a sensibilidade para estimativas dos parâmetros de regressão. Portanto, se a norma da matriz de variâncias-covariâncias estimada for muito grande, então, uma pequena perturbação na matriz de correlação de trabalho ou na resposta, poderá significar em um grande erro relativo para estimativas dos β 's.

Com isso, a proposta aqui é selecionar a matriz de correlação de trabalho espacial que minimiza Condition-Number-Based Criterion (CNBC) (3.15).

$$\begin{aligned} RJ_1M &= \frac{\|\hat{\mathbf{V}}_{rob}\| Cond(\hat{\mathbf{V}}_{Naive})}{p}, \\ RJ_2M &= \frac{(\|\hat{\mathbf{V}}_{rob}\| Cond(\hat{\mathbf{V}}_{Naive}))^2}{p}, \\ CNBC &= \sqrt{(RJ_1M - 1)^2 + (RJ_2M - 1)^2}, \end{aligned} \quad (3.15)$$

em que p é o número de covariáveis envolvidas no modelo marginal de regressão. Para avaliar o desempenho do critério *CNBC* é realizado um estudo de simulação, no qual os resultados são comparados com os critérios *QIC*, *CIC* e *RJC*, utilizando modelos de correlação gerados a partir das famílias Wendland, Matérn, e o modelo Esférico, obtidas pelas Equações (3.3.1), (3.16) e (3.17), respectivamente.

$$\rho(h) = \begin{cases} 1, & \text{se } h = 0 \\ \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{h}{\phi_3}\right)^{\nu} K_{\nu}\left(\frac{h}{\phi_3}\right), & \text{se } h > 0 \end{cases} \quad (3.16)$$

$$\rho(h) = \begin{cases} 1, & \text{se } h = 0 \\ 1 - \frac{3}{2} \frac{h}{\phi_3} + \frac{1}{2} \left(\frac{h}{\phi_3}\right)^3, & \text{se } h > 0 \end{cases} \quad (3.17)$$

em que h denota distância euclidiana entre duas localizações espaciais e, K_{ν} é a função de Bessel modificada do terceiro tipo de ordem.

3.4.1 Validação do critério de seleção de correlação espacial por estudo de simulação

Com objetivo de avaliar o desempenho do método proposto, realiza-se um estudo de simulação em termos da proporção de seleção de uma estrutura de correlação espacial verdadeira em comparação com os critérios *QIC*, *CIC* e *RJC*. A ideia é avaliar a taxa de acerto, ou

seja, o percentual em que os critérios identificou a verdadeira matriz de correlação que gerou os dados. Para gerar os dados utilizamos três cenários, a saber: considerando a variância da Normal, Poisson e Gama para dados espacialmente correlacionados.

Para gerar as respostas que tenham uma estrutura espacial e correlacionadas, utilizamos a função `mvrnorm()` do pacote *MASS* do *Software* (R Core Team, 2020) que gera valores da distribuição normal multivariada com a especificação de alguma estrutura de correlação. No caso das repostas da Poisson e Gama, utilizamos o algoritmo NORTA (*NORmal To Anything*), proposto inicialmente por Cario e Nelson (1997). O leitor interessado poderá consultar maiores detalhes sobre este algoritmo em Niaki e Abbasi (2008), Yahav e Shmueli (2012), Bonat (2017).

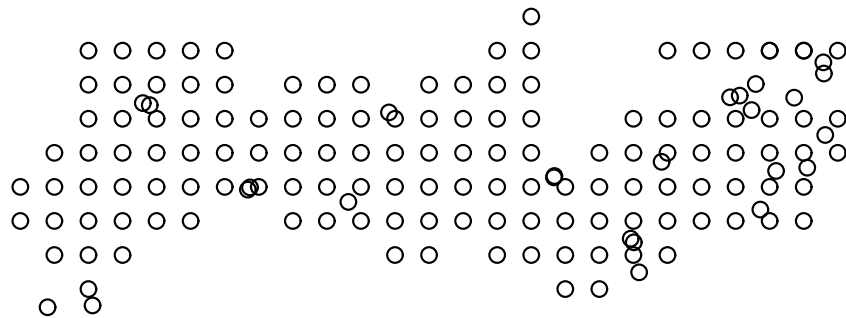
Para definir o modelo de regressão consideramos as funções de ligação identidade $g(\mu_i) = \mu_i$, logarítmica $g(\mu_i) = \log(\mu_i)$ e inversa $g(\mu_i) = \frac{1}{\mu_i}$ para distribuições Normal, Poisson e Gama, respectivamente. O parâmetro de regressão (β) é considerado fixo e igual a $\beta_0 = 0, 5$. As respostas são geradas sob estrutura de correlação espacial, considerando os modelos das famílias Wendland, Matérn e modelo Esférico. Consideramos uma suavização $\kappa = 0, 1, 2$ e 3 , e $\nu = 0, 5, 1, 5, 2, 5$ e $3, 5$ para covariâncias da família Wendland e Matérn, respectivamente. O vetor de parâmetros que caracteriza a estrutura espacial é considerado fixo para os três casos, igual a $\varphi = (\phi_1, \phi_2, \phi_3)^\top = (0, 2; 0, 6; 0, 9)^\top$. São consideradas três tipos de grades. Primeiramente consideramos uma grade de tamanhos 10×10 e 20×20 , Figuras 15(a) e 15(b), respectivamente, ambas com espaçamento de 1 (uma) unidade entre os pontos. Posteriormente, uma grade de um conjunto de dados reais sobre precipitação pluviométrica no primeiro semestre de 2021 no estado de Pernambuco, Brasil, coletados em 157 estações de medição, Figura 15.

Figura 14 – Grade de pontos com tamanhos 10×10 e 20×20 , Figuras 15(a) e 15(b), respectivamente, distribuídos com espaçamento de 1 (uma) unidade entre os pontos.



Fonte: Elaborada pelo autor (2022).

Figura 15 – Grade de pontos de um conjunto de dados reais sobre precipitação pluviométrica no primeiro semestre de 2021 no estado de Pernambuco, Brasil, referente a 157 estações e medição.



Fonte: Elaborada pelo autor (2022).

Definida a grade, em cada cenário é gerada 1000 réplicas de Monte Carlo, e em cada réplica é obtido um vetor de respostas $\mathbf{z} = (z_1, \dots, z_n)^\top$. Para tanto, as respostas e o ajuste do modelo de regressão sob EEG são realizados em três situações:

1. Define-se a verdadeira matriz de correlação espacial como sendo $W - \kappa$: Covariância Wendland com suavização κ , e gera as respostas. Então, ajusta o modelo sob EEG considerando a verdadeira matriz de correlação. Posteriormente, ajusta o modelo especificando uma estrutura de correlação $M - \nu$: Covariância da família Matérn com suavização ν , e com o modelo Esférico;

2. Define-se a verdadeira matriz de correlação espacial como sendo $M - \nu$: Covariância da família Matérn com suavização ν , e gera as respostas. Então, ajusta o modelo sob EEG considerando a verdadeira matriz de correlação. Posteriormente, ajusta o modelo especificando uma estrutura de correlação $W - \kappa$: Covariância da família Wendland com suavização κ , e com o modelo Esférico;
3. Define-se a verdadeira matriz de correlação espacial como sendo do modelo Esférico, e gera as respostas. Então, ajusta o modelo sob EEG considerando a verdadeira matriz de correlação. Posteriormente, ajusta o modelo especificando uma estrutura de correlação $W - \kappa$: Covariância da família Wendland com suavização κ , e com $M - \nu$: Covariância da família Matérn com suavização ν .

Com isso, para cada um dos casos acima é calculado os valores dos critérios QIC, CIC, RJC e CNBC. Então, é computada a taxa de acerto, dentre as 1000 réplicas, com relação a verdadeira correlação que gerou os dados, para cada critério. Os resultados podem ser vistos nas Tabelas 10, 11, 12, 13, 14 e 15.

Na Tabela 10, em que as respostas são normais e espacialmente distribuídos, notamos que quando a estrutura verdadeira de correlação é a Wendland, o critério QIC apresentou resultados ruins em todos os cenários, com relação ao taxa percentual de acertos da verdadeira matriz de correlação. Por outro lado, os critérios RJC e CNBC alcançaram ótimos resultados, nos quais obtiveram uma taxa de acerto de 100% em todos os cenário. Ainda, embora o critério CIC tenha obtido resultados razoáveis, suas taxas de acertos foram todas inferiores a 50%. Quando as respostas são geradas a partir da família Matérn, o CIC mantém os resultados ligeiramente equivalentes aos anteriores. No entanto, o QIC recuperou seu desempenho conseguindo ser altamente competitivo com os critérios RJC e CNBC. Neste caso, o QIC obteve uma taxa de acerto com variação no intervalo $[69, 5; 99, 8]$. Quando consideramos uma estrutura de correlação a partir do modelo Esférico, novamente o critério QIC apresentou péssimos resultados. Em contrapartida, critérios RJC e CNBC mostraram-se excelentes para seleção da verdadeira estrutura de correlação espacial.

Na Tabela 11, com respostas correlacionadas da Poisson, o critério CIC manteve seu desempenho sob razoabilidade. Por outro lado, o critério QIC apresentou competitivo com relação aos RJC e CNBC tanto para estrutura da Wendland quanto da Matérn. Destaca-se o desempenho do QIC, para estrutura da Matérn e parâmetro de suavização $\kappa = 3, 5$, em que alcançou o feito de uma taxa de acerto igual a 100%, superando o CNBC que alcançou uma

taxa igual a 98,7%. No entanto, para o modelo Esférico o Critério quasi-verossimilhança sob o modelo de independência (QIC) permaneceu com péssimos resultados. Na Tabela 12, com resposta correlacionada da Gama, os critérios Critério de informação de correlação (CIC) e QIC mantiveram desempenho razoáveis tanto para estrutura da Wendland quanto da Matérn e modelos Esférico, embora o QIC tenha obtido um desempenho muito ruim para o modelo Esférico considerando uma grade de tamanho 10×10 . Os critérios RJC e CNBC continuaram como excelentes resultados, com taxas de acertos superiores a 83% e 93%, respectivamente, em todas as configurações.

Quando a grade utilizada foi a dos dados reais, Tabelas 13, 14 e 15, os critérios mantiveram os mesmos padrões de desempenho que para uma grade simulada. Destarte, com ênfase em destacar o critério CNBC, aqui proposto, nota-se que o mesmo obteve excelentes resultados em todos os cenários, com taxa de acerto superior a 92%, alcançando 100% em alguns casos. O critério Rotnizky and Jewell's criterion (RJC) alcançou uma taxa superior a 80%. Portanto, tanto o critério aqui proposto, quanto o RJC podem ser considerados para auxiliar na escolha da matriz de correlação espacial. Ademais, convém ressaltar que o critério Condition-Number-Based Criterion (CNBC) leva em consideração os efeitos dos erros de arredondamento envolvidos no processo de inversão matricial, que estão presentes nos estimadores robusto e *naive*. Contudo, vale frisar que a proposta está pautada sob a suposição de que o modelo de média marginal está especificado corretamente. Sendo assim, é interessante que se realize investigações futuras para avaliar o desempenho do critério na seleção de covariáveis no preditor linear, sob a suposição da especificação correta da matriz de correlação espacial. Com relação ao critério QIC, suas limitações em selecionar a verdadeira estrutura de correlação já haviam sido destacadas nos trabalhos de Hin, Carey e Wang (2007), Hin e Wang (2009), Gosho, Hamada e Yoshimura (2011), Jaman et al. (2016) e Silva e Cirillo (2018). Com isso, os seus resultados ruins vão de acordo com os apresentados na literatura com relação a identificação da verdadeira matriz de correlação.

Tabela 10 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Normal. Os pontos são distribuídos em grades regulares de tamanho 10×10 e 20×20 . Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.

Estrutura Verdadeira	Critério	Especificação com $n = 100$			Especificação com $n = 400$		
		W - κ	M - 4,5	Esf	W - κ	M - 4,5	Esf
W - 0	CIC	47,5	13,3	39,2	49,6	4,6	45,8
	QIC	0,0	99,9	0,1	2,7	95,5	1,8
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
W - 1	CIC	35,2	18,5	46,3	35,4	6,1	58,5
	QIC	0,0	100	0,0	0,6	96,8	2,6
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
W - 2	CIC	36,6	16,4	47,0	35,7	7,1	57,2
	QIC	0,0	99,7	0,0	0,3	97,0	2,7
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
W - 3	CIC	38,1	17,4	44,5	38,0	4,9	57,1
	QIC	0,1	99,9	0,0	0,2	97,7	2,1
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
M - 0,5		M - ν	W - 3	Esf	M - ν	W - 3	Esf
	CIC	28,8	27,5	43,7	34,2	25,3	40,5
	QIC	91,6	1,9	6,5	79,7	7,8	12,5
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
M - 1,5	CIC	21,6	32,3	46,1	20,5	32,8	46,7
	QIC	99,6	0,0	0,4	69,5	2,2	28,3
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
M - 2,5	CIC	22,1	32,8	45,1	19,0	34,2	46,1
	QIC	99,3	0,1	0,6	90,5	1,0	8,5
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
M - 3,5	CIC	21,2	32,6	46,2	18,3	34,2	47,5
	QIC	99,8	0,1	0,1	93,0	0,6	6,4
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0
Esférico		Esf	W - 3	M - 4,5	Esf	W - 3	M - 4,5
	CIC	50,2	33,9	15,9	59,0	34,0	7,0
	QIC	0,1	0,0	99,9	3,9	0,2	95,9
	RJC	100	0,0	0,0	100	0,0	0,0
	CNBC	100	0,0	0,0	100	0,0	0,0

Fonte: Elaborada pelo autor (2022).

CNBC: Condition-Number-Based Criterion. M - ν : Matérn com suavização ν . W - κ : Wendland com suavização κ .

Tabela 11 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Poisson. Os pontos são distribuídos em grades regulares de tamanho 10×10 e 20×20 . Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.

Estrutura Verdadeira	Critério	Especificação com $n = 100$			Especificação com $n = 400$		
		W - κ	M - 4,5	Esf	W - κ	M - 4,5	Esf
W - 0	CIC	48,1	31,6	20,3	45,1	36,1	18,8
	QIC	70,4	14,8	14,8	1,5	97,7	0,8
	RJC	91,6	1,6	6,8	91,5	5,3	3,2
	CNBC	100	0,0	0,0	100	0,0	0,0
W - 1	CIC	42,8	26,2	31,0	38,7	16,7	44,6
	QIC	68,4	11,1	20,5	0,5	98,0	1,5
	RJC	90,0	2,5	7,5	91,8	5,0	3,2
	CNBC	100	0,0	0,0	99,8	0,2	0,0
W - 2	CIC	40,1	26,6	33,3	45,0	13,6	41,4
	QIC	73,7	10,3	16,0	0,0	99,4	0,6
	RJC	92,2	1,9	5,9	92,4	4,5	3,1
	CNBC	100	0,0	0,0	99,9	0,1	0,0
W - 3	CIC	38,3	25,1	36,6	40,9	17,7	41,4
	QIC	74,9	12,3	12,8	0,0	99,2	0,8
	RJC	91,3	2,1	6,6	89,5	5,9	4,6
	CNBC	100	0,0	0,0	100	0,0	0,0
M - 0,5		M - ν	W - 3	Esf	M - ν	W - 3	Esf
	CIC	43,2	38,5	18,3	39,9	37,3	22,8
	QIC	92,9	2,6	4,5	78,4	8,3	13,3
	RJC	93,3	3,5	3,2	91,7	5,5	2,8
	CNBC	99,0	1,0	0,0	99,5	0,5	0,0
M - 1,5	CIC	18,2	30,9	50,9	39,2	30,3	30,5
	QIC	99,1	0,9	0,0	67,0	3,3	29,7
	RJC	89,8	5,1	5,1	92,3	5,4	2,3
	CNBC	100	0,0	0,0	99,5	0,2	0,3
M - 2,5	CIC	23,3	33,0	43,7	34,4	33,7	31,9
	QIC	99,7	0,0	0,3	90,4	0,7	8,9
	RJC	92,4	3,9	3,7	91,0	5,6	3,4
	CNBC	98,8	1,2	0,0	99,2	0,8	0,0
M - 3,5	CIC	36,9	32,2	30,9	20,5	34,8	44,7
	QIC	100	0,0	0,0	92,4	0,4	7,2
	RJC	89,3	5,0	5,7	92,4	4,0	3,6
	CNBC	98,7	1,3	0,0	99,2	0,8	0,0
Esférico		Esf	W - 3	M - 4,5	Esf	W - 3	M - 4,5
	CIC	30,5	38,2	31,3	41,0	30,5	28,5
	QIC	0,1	0,0	99,9	5,7	0,1	94,2
	RJC	94,5	4,2	1,3	94,9	3,9	1,2
	CNBC	99,8	0,0	0,2	100	0,0	0,0

Fonte: Elaborada pelo autor (2022).

CNBC: Condition-Number-Based Criterion. M - ν : Matérn com suavização ν . W - κ : Wendland com suavização κ .

Tabela 12 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Gama. Os pontos são distribuídos em grades regulares de tamanho 10×10 e 20×20 . Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.

Estrutura Verdadeira	Critério	Especificação com $n = 100$			Especificação com $n = 400$		
		W - κ	M - 4,5	Esf	W - κ	M - 4,5	Esf
W - 0	CIC	19,4	40,0	40,6	38,4	40,1	21,5
	QIC	37,2	56,0	6,8	8,8	53,2	38,0
	RJC	95,0	2,6	2,4	86,5	5,4	8,1
	CNBC	97,0	1,4	1,6	95,2	2,1	2,7
W - 1	CIC	25,9	35,8	38,3	32,0	31,2	36,8
	QIC	42,1	49,5	8,4	37,4	54,0	8,6
	RJC	89,4	2,7	7,9	84,5	8,5	7,0
	CNBC	95,7	1,8	2,5	93,4	4,0	2,6
W - 2	CIC	34,1	32,8	33,1	35,5	29,7	34,8
	QIC	41,5	51,3	7,2	36,7	55,4	7,9
	RJC	89,2	2,4	8,4	85,3	8,8	5,9
	CNBC	95,6	2,1	2,3	93,7	3,9	2,4
W - 3	CIC	39,6	30,0	30,4	32,5	33,1	34,4
	QIC	43,7	52,7	3,6	35,8	57,0	7,2
	RJC	89,5	3,1	7,4	88,4	6,6	5,5
	CNBC	95,5	2,0	2,5	95,1	3,7	1,2
M - 0,5		M - ν	W - 3	Esf	M - ν	W - 3	Esf
	CIC	34,4	38,3	27,3	40,2	36,6	23,2
	QIC	41,0	54,0	5,0	35,8	47,3	16,9
	RJC	83,6	8,7	7,7	84,6	9,0	6,4
M - 1,5	CNBC	94,2	3,1	2,7	94,9	3,3	1,8
	CIC	23,7	34,7	41,6	32,7	31,1	36,2
	QIC	43,0	50,7	6,3	35,4	37,5	27,1
	RJC	87,6	7,5	4,9	84,2	9,4	6,4
M - 2,5	CNBC	95,1	3,2	1,7	94,0	3,9	2,1
	CIC	25,4	39,9	34,7	28,0	32,7	39,3
	QIC	42,5	52,3	5,2	38,9	42,5	18,6
	RJC	86,7	6,2	7,1	84,3	8,4	7,3
M - 3,5	CNBC	95,1	2,3	2,6	93,4	3,7	2,9
	CIC	30,2	37,7	32,1	32,1	28,9	39,0
	QIC	44,0	49,5	6,5	40,4	41,2	18,4
	RJC	85,7	6,6	7,7	89,2	5,5	5,3
Esférico	CNBC	94,7	2,7	2,6	96,1	1,8	2,1
		Esf	W - 3	M - 4,5	Esf	W - 3	M - 4,5
	CIC	31,0	43,3	25,7	35,9	40,8	23,3
	QIC	3,3	49,7	47,0	39,9	9,1	51,0
	RJC	90,4	7,6	2,0	89,3	5,1	5,6
	CNBC	96,0	2,7	1,3	95,4	1,6	3,0

Fonte: Elaborada pelo autor (2022).

CNBC: Condition-Number-Based Criterion. M - ν : Matérn com suavização κ . W - κ : Wendland com suavização ν .

Tabela 13 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Normal. Os pontos são distribuídos em uma grade de um conjunto de dados reais sobre precipitação no estado de Pernambuco, Brasil. Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.

Estrutura Verdadeira	Critério	W - κ	M - 4, 5	Esf
W - 0	<i>CIC</i>	37,3	26,1	36,6
	<i>QIC</i>	0,5	99,4	0,1
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0
W - 1	<i>CIC</i>	33,2	28,3	38,5
	<i>QIC</i>	0,1	99,4	0,5
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0
W - 2	<i>CIC</i>	33,2	29,6	37,2
	<i>QIC</i>	0,0	99,9	0,1
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0
W - 3	<i>CIC</i>	31,9	28,6	39,5
	<i>QIC</i>	0,3	99,4	0,3
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0
M - 0,5		M - ν	W - 3	Esf
	<i>CIC</i>	35,5	30,3	34,2
	<i>QIC</i>	82,8	7,7	9,5
	<i>RJC</i>	100	0,0	0,0
M - 1,5	<i>CIC</i>	30,1	30,6	39,3
	<i>QIC</i>	98,0	0,4	1,6
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0
M - 2,5	<i>CIC</i>	30,4	32,4	37,2
	<i>QIC</i>	98,7	0,4	0,9
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0
M - 3,5	<i>CIC</i>	26,9	33,6	39,5
	<i>QIC</i>	99,8	0,0	0,2
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0
Esférico		Esf	W - 3	M - 4, 5
	<i>CIC</i>	39,4	31,5	29,1
	<i>QIC</i>	0,5	0,1	99,4
	<i>RJC</i>	100	0,0	0,0
	<i>CNBC</i>	100	0,0	0,0

Fonte: Elaborada pelo autor (2022).

CNBC: Condition-Number-Based Criterion. M - ν : Matérn com suavização ν . W - κ : Wendland com suavização κ .

Tabela 14 – Proporções (%) de acertos na seleção da estrutura de correlação espacial para respostas correlacionada da Poisson. Os pontos são distribuídos em uma grade de um conjunto de dados reais sobre precipitação no estado de Pernambuco, Brasil. Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.

Estrutura Verdadeira	Critério	W - κ	M - 4, 5	Esf
W - 0	<i>CIC</i>	24,5	56,5	19,0
	<i>QIC</i>	0,1	99,8	0,1
	<i>RJC</i>	93,8	3,3	2,9
	<i>CNBC</i>	99,7	0,3	0,0
W - 1	<i>CIC</i>	18,2	54,8	27,0
	<i>QIC</i>	0,0	99,8	0,2
	<i>RJC</i>	90,4	5,6	4,0
	<i>CNBC</i>	98,8	1,2	0,0
W - 2	<i>CIC</i>	20,4	52,5	27,1
	<i>QIC</i>	0,0	99,8	0,2
	<i>RJC</i>	90,1	4,4	5,5
	<i>CNBC</i>	99,8	0,2	0,0
W - 3	<i>CIC</i>	20,3	52,9	26,8
	<i>QIC</i>	0,1	99,5	0,4
	<i>RJC</i>	90,3	5,4	4,3
	<i>CNBC</i>	99,4	0,6	0,0
M - 0,5		M - ν	W - 3	Esf
	<i>CIC</i>	51,9	23,3	24,8
	<i>QIC</i>	74,9	11,9	13,2
	<i>RJC</i>	94,5	1,8	3,7
M - 1,5	<i>CIC</i>	23,9	39,8	36,3
	<i>QIC</i>	98,6	0,0	1,4
	<i>RJC</i>	90,8	3,6	5,6
	<i>CNBC</i>	100	0,0	0,0
M - 2,5	<i>CIC</i>	53,9	16,3	29,8
	<i>QIC</i>	99,6	0,0	0,4
	<i>RJC</i>	95,5	2,9	3,6
	<i>CNBC</i>	100	0,0	0,0
M - 3,5	<i>CIC</i>	51,6	18,8	29,6
	<i>QIC</i>	99,2	0,1	0,7
	<i>RJC</i>	93,3	2,8	3,9
	<i>CNBC</i>	99,1	0,9	0,0
Esférico		Esf	W - 3	M - 4, 5
	<i>CIC</i>	26,2	23,1	50,7
	<i>QIC</i>	0,3	0,0	99,7
	<i>RJC</i>	93,7	3,7	2,6
	<i>CNBC</i>	98,9	0,7	0,4

Fonte: Elaborada pelo autor (2022).

CNBC: Condition-Number-Based Criterion. M - ν : Matérn com suavização ν . W - κ : Wendland com suavização κ .

Tabela 15 – Proporções (%) de certos na seleção da estrutura de correlação espacial para respostas correlacionada da Gama. Os pontos são distribuídos em uma grade de um conjunto de dados reais sobre precipitação no estado de Pernambuco, Brasil. Os dados são gerados alternando os modelos da família Wendland, Matérn e Esférico.

Estrutura Verdadeira	Critério	W - κ	M - 4, 5	Esf
W - 0	<i>CIC</i>	16,8	50,6	32,6
	<i>QIC</i>	39,1	52,7	8,2
	<i>RJC</i>	91,5	5,7	2,8
	<i>CNBC</i>	95,1	3,4	1,5
W - 1	<i>CIC</i>	22,8	51,3	25,9
	<i>QIC</i>	30,2	44,4	25,4
	<i>RJC</i>	87,0	8,0	5,0
	<i>CNBC</i>	93,3	4,7	2,0
W - 2	<i>CIC</i>	23,4	50,5	26,1
	<i>QIC</i>	34,2	44,4	21,4
	<i>RJC</i>	87,2	9,3	3,5
	<i>CNBC</i>	92,6	5,7	1,7
W - 3	<i>CIC</i>	27,6	45,2	27,2
	<i>QIC</i>	40,1	44,6	15,3
	<i>RJC</i>	85,5	9,5	5,0
	<i>CNBC</i>	93,5	4,2	2,3
M - 0,5		M - ν	W - 3	Esf
	<i>CIC</i>	45,3	35,5	19,2
	<i>QIC</i>	37,1	47,8	15,1
	<i>RJC</i>	84,7	8,3	7,0
M - 1,5	<i>CNBC</i>	94,6	3,9	1,5
	<i>CIC</i>	25,2	29,7	45,1
	<i>QIC</i>	40,8	41,8	17,4
	<i>RJC</i>	86,3	7,6	6,1
M - 2,5	<i>CNBC</i>	95,0	2,5	2,5
	<i>CIC</i>	46,2	23,2	30,6
	<i>QIC</i>	40,9	41,8	17,3
	<i>RJC</i>	88,1	6,3	5,6
M - 3,5	<i>CNBC</i>	95,7	2,1	2,2
	<i>CIC</i>	48,1	25,0	26,9
	<i>QIC</i>	41,4	43,7	14,9
	<i>RJC</i>	89,2	6,6	4,2
Esférico	<i>CNBC</i>	97,2	1,5	1,3
		Esf	W - 3	M - 4, 5
	<i>CIC</i>	30,0	32,6	37,4
	<i>QIC</i>	10,3	42,4	47,3
	<i>RJC</i>	87,1	7,5	5,4
	<i>CNBC</i>	95,1	2,1	2,8

Fonte: Elaborada pelo autor (2022).

CNBC: Condition-Number-Based Criterion. M - ν : Matérn com suavização ν . W - κ : Wendland com suavização κ .

3.5 MEDIDAS DE INFLUÊNCIA LOCAL

Nesta Seção os conceitos de influência local propostos por Cook (1986) são aplicados no estudo de modelos de quasi-verossimilhança para dados correlacionados. A metodologia utilizada consiste de uma extensão dos trabalhos: Jung (2007), em que é desenvolvido um método de diagnóstico para os estimadores de quasi-verossimilhança em modelos lineares generalizados, Jung (2008) que utilizam o deslocamento de quasi-probabilidade e propõem técnicas de diagnósticos de influência local sob vários esquemas de perturbação. Venezuela, Sandoval e Botter (2011) que estudam medidas influência local baseados em equações de estimação. Nava et al. (2017) na qual aplica a técnica de influência local em um modelo espacial com distribuição marginal binomial e função de ligação logit. Embora estes trabalhos tenham fornecidos importantes abordagens metodológicas, sobre os vários caminhos possíveis de se estudar e/ou aplicar as extensões das medidas de influência local, ainda que a última autora tenha estudado tais medidas para um modelo espacial com distribuição marginal binomial, não foram abordadas a metodologia no contexto geral para dados espacialmente correlacionados. Diante disso, a proposta aqui é realizar esta extensão e utilizar como matriz de correlação espacial de trabalho um membro da família de Wendland.

Assim, por analogia a influência local pelo afastamento das verossimilhanças (ver COOK) define-se uma medida de afastamento das quasi-verossimilhanças para dados correlacionados por (3.18).

$$QD(\omega) = 2 \left(Q(\hat{\beta}; \mathbf{Z}) - Q(\hat{\beta}_{\omega}; \mathbf{Z}) \right), \quad (3.18)$$

em que $\hat{\beta}$ e $\hat{\beta}_{\omega}$ denotam os estimadores de regressão por EEG para os modelos não-perturbados e perturbados, respectivamente, para β , com em que $\omega = (\omega_1, \dots, \omega_n)^T$ um vetor de perturbação de dimensão $n \times 1$ em $\Omega \in \mathbb{R}^n$ denotando um conjunto de perturbações. Ainda, assume-se que $Q(\beta|\omega_0) = Q(\beta)$, com ω_0 um vetor de não perturbação.

Desta maneira, o comportamento local do deslocamento das quasi - verossimilhanças é mensurado a partir da curvatura normal na direção de algum vetor unitário. Então, a curvatura normal C_1 na direção de um vetor unitário $\mathbf{1}$ em ω_0 é $C_1 = 2|\mathbf{1}^T \ddot{\mathbf{F}} \mathbf{1}|$, em que $\ddot{\mathbf{F}} = -\frac{\partial^2 Q(\hat{\beta}_{\omega})}{\partial \omega \partial \omega^T} \Big|_{\omega=\omega_0} = -\Delta^T \ddot{\mathbf{Q}}_{\beta\beta}^{-1} \Delta$, com $\Delta = \frac{\partial^2 Q(\hat{\beta}|\omega)}{\partial \beta \partial \omega^T} \Big|_{\beta=\hat{\beta}; \omega=\hat{\omega}_0}$ representando

uma matriz de ordem $p \times n$ e, $\ddot{Q}_{\beta\beta}(\hat{\beta}) = \frac{\partial^2 Q(\beta)}{\partial \beta \partial \beta^\top} \Big|_{\beta=\hat{\beta}}$ é a matriz Hessiana de ordem $p \times p$

com elementos dados por:

$$\frac{\partial^2 Q(\beta; \mathbf{Z})}{\partial \beta \partial \beta_j} = \mathbf{X}^\top \left\{ \frac{\partial (\dot{\mathbf{K}} \mathbf{V}^{-1})}{\partial \beta_j} (\mathbf{Z} - \boldsymbol{\mu}) - \dot{\mathbf{K}} \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \beta_j} \right\} \quad \text{para } j = 1, \dots, p.$$

Estes resultados decorrem de Jung (2007) e Nava et al. (2017) nos quais podem ser consultados maiores detalhes. Os elementos para as derivadas parciais de $\dot{\mathbf{K}} = \text{diag} \left[\frac{\partial \mu_i}{\partial \eta_i} \right]$ e $\boldsymbol{\mu}$ com relação a β_j são dados por $\frac{\partial^2 \mu_i}{\partial \eta_i^2} \frac{\partial \boldsymbol{\eta}}{\partial \beta_j}$ e $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \beta_j}$, respectivamente, as quais dependem da função de ligação adotada. No trabalho de Nava et al. (2017) são apresentados os resultados para distribuição marginal Binomial e função de ligação logit. Neste trabalho, como contribuições para literatura os resultados por ora mencionados são estendidos, substituindo a matriz de correlação de trabalho por uma pertencente a família Wendland. Além disso, são desenvolvidos os resultados para outras distribuições marginais, os quais estão apresentados na Tabela 16, com detalhes no Apêndice D. As expressões para os elementos da matriz Δ , aqui são apresentados segundo os cenários de perturbação de na variável resposta, no preditor linear e na matriz de correlação espacial de trabalho.

Tabela 16 – Funções de ligação e derivadas da média (μ) utilizadas na expressão dos elementos da matriz Hessiana das medidas de influência local para várias distribuições da família exponencial.

Distribuição	Função de ligação	$\frac{\partial \mu_i}{\partial \eta_i}$	$\frac{\partial \dot{\mathbf{K}}}{\partial \beta_j}$	$V(\mu)$
Binomial	$g(\mu) = \log\left(\frac{\pi}{1-\pi}\right)$	$\pi_i(1 - \pi_i)$	$\pi_i(1 - \pi_i)(1 - 2\pi_i)x_{ij}$	$\pi_i(1 - \pi_i)$
Poisson	$g(\mu) = \log(\mu)$	$\exp(x_{ij}\beta_j)$	$\exp(x_{ij}\beta_j)x_{ij}$	μ_i
Binomial Negativa	$g(\mu) = \log(\mu)$	$\exp(x_{ij}\beta_j)$	$\exp(x_{ij}\beta_j)x_{ij}$	$\mu_i + \frac{\mu_i}{k}$
Gama	$g(\mu) = \frac{1}{\mu}$	$-\frac{1}{(x_{ij}\beta_j)^2}$	$\frac{2x_{ij}}{(x_{ij}\beta_j)^3}$	μ_i^2
Normal Inversa	$g(\mu) = \frac{1}{\mu^2}$	$\frac{1}{2}(x_{ij}\beta_j)^{-1/2}$	$-\frac{1}{4}(x_{ij}\beta_j)^{-3/2}x_{ij}$	μ_i^3

Fonte: Elaborada pelo autor (2022).

3.5.1 Esquemas de perturbação

Com o objetivo de identificar observações potencialmente influentes, realiza-se uma ponderação à contribuição do afastamento da quasi - verossimilhança. Sendo assim, são considerados os esquemas de perturbação já bem conhecidos na literatura: na variável resposta, matriz de correlação de trabalho e no preditor linear. Ademais, aqui, embasado nestas ideias, sugere-se

uma perturbação aplicada direta no vetor de parâmetros que define a estrutura de dependência espacial $\boldsymbol{\varphi} = (\phi_1, \phi_2, \phi_3)^\top$.

3.5.1.1 Perturbação na variável resposta e no preditor linear

Para uma perturbação na variável resposta $(z_1, \dots, z_n)^\top$, considerando uma perturbação do tipo aditiva substitui-se z_i , por $z_{\omega_i} = z_i + \omega_i \sqrt{\text{var}(z_i)}$, $i = 1, \dots, n$. Então, a equação de quasi-verossimilhança perturbada é escrita como em (3.19).

$$\frac{\partial Q(\boldsymbol{\beta}; \boldsymbol{\omega})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \beta_j} \right)^\top \mathbf{V}^{-1} (\mathbf{Z}_\omega - \boldsymbol{\mu}). \quad (3.19)$$

Logo, a matriz Δ é dada por:

$$\Delta = \mathbf{X}^\top \dot{\mathbf{K}} \mathbf{V}^{-1} \frac{\partial (\mathbf{Z}_\omega - \boldsymbol{\mu})}{\partial \omega_i^\top} = \mathbf{X}^\top \dot{\mathbf{K}} \mathbf{V}^{-1} \mathbf{E},$$

em que $\mathbf{E} = \text{diag} [\sqrt{\text{var}(z_i)}]$, com $i = 1, \dots, n$.

No caso da perturbação no preditor linear, substitui-se $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ por $\boldsymbol{\eta}_\omega = \boldsymbol{\eta} + \boldsymbol{\omega}$. Então, quasi-verossimilhança perturbada é escrita como:

$$\frac{\partial Q(\boldsymbol{\beta}; \boldsymbol{\omega})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \boldsymbol{\mu}_\omega}{\partial \beta_j} \right)^\top \mathbf{V}_\omega^{-1} (\mathbf{Z} - \boldsymbol{\mu}_\omega),$$

em que $\frac{\partial \mu_{\omega_i}}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_{\omega_i}} \frac{\partial \eta_{\omega_i}}{\partial \beta_j}$, com $\frac{\partial \eta_{\omega_i}}{\partial \beta_j} = x_{ij}$. Portanto,

$$\Delta = \mathbf{X}^\top \left\{ \frac{\partial \dot{\mathbf{K}} \boldsymbol{\omega} \mathbf{V}_\omega^{-1}}{\partial \omega_i} (\mathbf{Z} - \boldsymbol{\mu}_\omega) - \dot{\mathbf{K}} \boldsymbol{\omega} \mathbf{V}_\omega^{-1} \frac{\partial \boldsymbol{\mu}_\omega}{\partial \omega_i} \right\}.$$

3.5.1.2 Perturbação na matriz de correlação de trabalho

Seja $\varrho_{ij} = 1 - \gamma(h_{ij}, \boldsymbol{\varphi})$, as entradas de uma matriz de correlação da família Wendland, com $\gamma(h_{ij}, \boldsymbol{\varphi})$ definindo uma função de semivariância da família Wendland. Em Venezuela, Sandoval e Botter (2011) é sugerido que seja perturbado o vetor de correlação, aqui, o vetor que mensura a estrutura de dependência espacial entre as observações. Os resultados aqui apresentados são distintos dos obtidos pela autora ora citada, haja vista que a estrutura da matriz de correlação consideradas não apresentam similaridades. Neste caso, tem-se que $\phi_{\omega_i} = \frac{\phi_j}{\omega_i}$, com $j = 1, \dots, 3$ e $i = 1, \dots, n$. Logo, a equação de quase-verossimilhança perturbada é escrita como em (3.20).

$$\frac{\partial Q(\boldsymbol{\beta}; \boldsymbol{\omega})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \beta_j} \right)^\top \mathbf{V}_\omega^{-1} (\mathbf{Z} - \boldsymbol{\mu}). \quad (3.20)$$

Neste caso, a expressão para matriz Δ é:

$$\Delta = \mathbf{X}^\top \dot{\mathbf{K}} \frac{\partial \mathbf{V}_{\omega}^{-1}}{\partial \omega^\top} (\mathbf{Z} - \boldsymbol{\mu}),$$

em que $\frac{\partial \mathbf{V}_{\omega}^{-1}}{\partial \omega^\top} = -\mathbf{V}_{\omega}^{-1} \frac{\partial \mathbf{V}_{\omega}}{\partial \omega^\top} \mathbf{V}_{\omega}^{-1}$ e $\frac{\partial \mathbf{V}_{\omega}}{\partial \omega^\top} = v^{1/2} \frac{\partial \mathbf{R}(\varphi/\omega)}{\partial \omega^\top} v^{1/2}$. Ainda,

$$\frac{\partial \mathbf{R}(\varphi/\omega)}{\partial \omega^\top} = \frac{\partial}{\partial \omega_i^\top} \begin{cases} \frac{\phi_2}{\omega_i} \rho_{\mu, \kappa} \left(\frac{\omega_i h_{ij}}{\phi_3} \right), & \text{se } i \neq j \\ \frac{1}{\omega_i} (\phi_1 + \phi_2), & \text{se } i = j \end{cases} \quad (3.21)$$

em que $\frac{\partial \rho_{\mu, \kappa}(\frac{\omega_i h}{\phi_3})}{\partial \omega_i^\top}$ é avaliada em $\omega = \omega_0$. As expressões para vários valores de κ e μ estão no Apêndice (E).

Dessarte, destaca-se que o resultado obtido em (3.21) poderá ser aplicado para qualquer tipo de função de correlação espacial válida, bastando apenas substituir a função $\rho(\cdot)$ pelas correspondente, por exemplo, família Matérn. Dito isto, na Tabela 17 estes resultados são aplicados à distribuição marginal da Binomial, Poisson, Binomial Negativa, Gama e Normal Inversa.

Tabela 17 – Resultados para os esquemas de perturbação

Distribuição	μ_{ω_i}	$\frac{\partial \mu_{\omega}}{\partial \beta_j}$	$\dot{\mathbf{K}}_{\omega}$	$\frac{\partial \dot{\mathbf{K}}_{\omega}}{\partial \omega_i}$
Binomial	$m_i \frac{\exp(\eta_i + \omega_i)}{1 + \exp(\eta_i + \omega_i)}$	$m_i \pi_{\omega_i} (1 - \pi_{\omega_i}) x_{ij}$	$\text{diag} [m_i \pi_{\omega_i} (1 - \pi_{\omega_i})]$	$\text{diag} [m_i \pi_{\omega_i} (1 - \pi_{\omega_i}) (1 - 2\pi_{\omega_i}) \sqrt{\text{var}(z_i)}]$
Poisson	$\exp(\eta_i + \omega_i)$	$\exp(\eta_i + \omega_i) x_{ij}$	$\text{diag} [\exp(\eta_i + \omega_i)]$	$\text{diag} [\exp(\eta_i + \omega_i) \sqrt{\text{var}(z_i)}]$
B. Negativa	$m_i \exp(\eta_i + \omega_i)$	$m_i \exp(\eta_i + \omega_i) x_{ij}$	$\text{diag} \left[-\frac{1}{(\eta_i + \omega_i)^2} \right]$	$\text{diag} [m_i \exp(\eta_i + \omega_i) \mathbf{e}_i]$
Gama	$\frac{1}{\eta_i + \omega_i}$	$-\frac{1}{(\eta_i + \omega_i)^2} x_{ij}$	$\text{diag} \left[-\frac{1}{(\eta_i + \omega_i)^2} \right]$	$\text{diag} \left[\frac{2}{(\eta_i + \omega_i)^3} \sqrt{\text{var}(z_i)} \right]$
Normal Inversa	$\sqrt{\frac{1}{\eta_i + \omega_i}}$	$-\frac{1}{2} \left(\frac{1}{\eta_i + \omega_i} \right)^{5/2} x_{ij}$	$\text{diag} \left[-\frac{1}{2} \left(\frac{1}{\eta_i + \omega_i} \right)^{5/2} \right]$	$\text{diag} \left[\frac{5}{4} \left(\frac{1}{\eta_i + \omega_i} \right)^{7/2} \sqrt{\text{var}(z_i)} \right]$

Fonte: Elaborada pelo autor (2022).

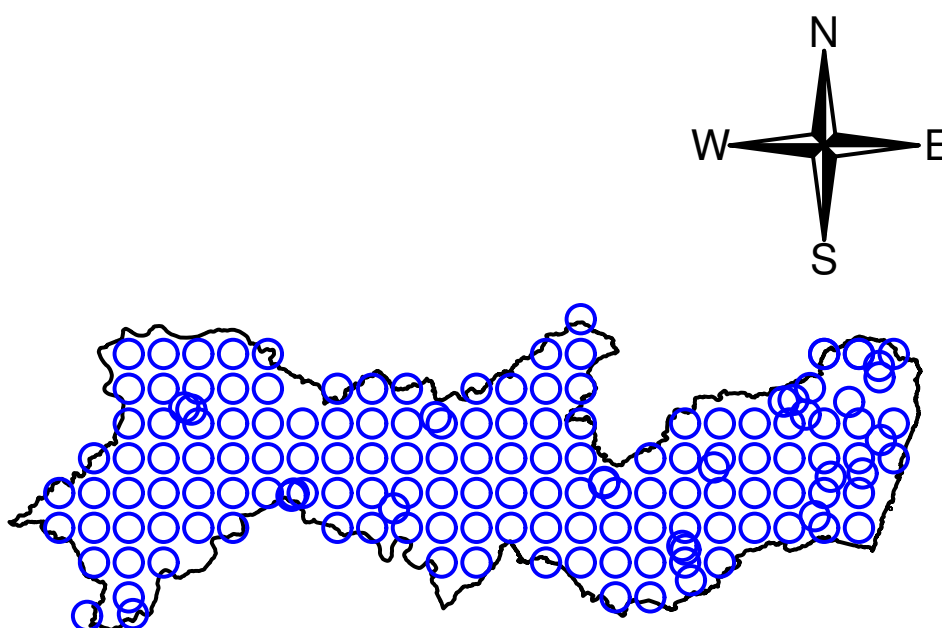
3.6 APLICAÇÃO: ANÁLISE ESPACIAL DA PRECIPITAÇÃO PLUVIOMÉTRICA (MM) PARA O PRIMEIRO SEMESTRE DE 2021 NO ESTADO DE PERNAMBUCO

Como é bem discutido na literatura, é crescente a necessidade de estudos relacionados a modelagem espacial da quantidade de chuva em determinada região, tendo a geoestatística como poderosa ferramenta para este fim. A problemática é vista como de extrema importância, principalmente em regiões agrícolas e metropolitanas, pois a ocorrência de chuvas em excesso ou sua escassez tem como consequência potenciais danos socioeconômicos e/ou ambientais, como exemplo, o abastecimento urbano, geração de energia elétricas, focos de incêndios,

influência no rendimento de culturas agrícolas, dentre outros. Diante disso, este estudo tem como objetivo examinar a aplicabilidade da metodologia aqui desenvolvida, com vistas a avaliar a distribuição espacial da precipitação pluviométrica (mm) média, para o primeiro semestre de 2021, no Estado de Pernambuco, Brasil (BR).

O Estado de Pernambuco está localizado na região Nordeste do Brasil, com área territorial, segundo dados do IBGE (2020) de $98.067,880\text{km}^2$ distribuída em 184 municípios, com uma população estimada de 9.616.621 em 2020. Com isso, para realizar este estudo foram coletados dados da precipitação pluviométrica a partir de 157 postos pluviométricos, representando todo o Estado. No entanto, 3 das estações foram excluídas do conjunto de dados, pois apresentaram localizações fora dos limites do estado. Os dados de precipitação utilizados foram coletados diretamente do banco de dados online do Sistema de Monitoramento Agrometeorológico (Agritempo), disponível em <https://www.agritempo.gov.br/agritempo/index.jsp>, sendo calculadas as médias de precipitação entre os meses de Janeiro e Junho de 2021 para cada estação pluviométrica. Os pontos geográficos estão apresentados na Figura 16.

Figura 16 – Mapa da localização espacial de 154 estações meteorológicas no Estado de Pernambuco.



Fonte: Elaborada pelo autor (2022).

Então, para entender o comportamento dos dados, inicialmente, realiza-se sua análise exploratória, de modo que seja possível verificar se os mesmos representam informações adequadas para ajustes de modelos sob dados espacialmente correlacionados. Um estudo descritivo é realizado com as principais medidas listadas na Tabela 18. Como podemos verificar, os da-

dos de precipitação apresentam alta heterogeneidade vista pelos coeficiente de variação. Esta heterogeneidade é característica devido ao fatos das estações localizadas no sertão e litoral do estado, que por apresentarem características climáticas peculiares, em geral não seguem o mesmo padrão de distribuição dos demais locais. Com vistas nos coeficiente de assimetria e curtose, percebe-se que os dados não podem ser classificados como normalmente distribuídos, isto é, assimetria $\neq 0$ e curtose $\neq 3$. Esta afirmativa, também, pode ser verificada pelo histograma, Figura 18(a), no evidencia uma leve assimetria já destacada pelo coeficiente de assimetria (0,76). No *boxplot*, Figura 18(b), é destacado quatro observações com valores altos que podem ser considerados como atípicos. Ainda, com vistas na Figura 19(a), gráfico de *postplot* classificados segundo os quartis, notamos que a região oeste do estado (sertão) apresentou altas quantidades de chuva, para o primeiro semestre de 2021. Embora, em poucas estações, a região litorânea e entorno apresentaram alguns valores altos e medianos. Sendo percebido uma quantidade considerável de medidas altas no norte do estado. A Figura 19(b), que representa o gráfico de semivariograma com envelope simulado, destaca a presença de alguns pontos fora do envelope, indicando uma considerável dependência espacial, isto é, rejeita-se a hipótese de independência espacial entre as observações.

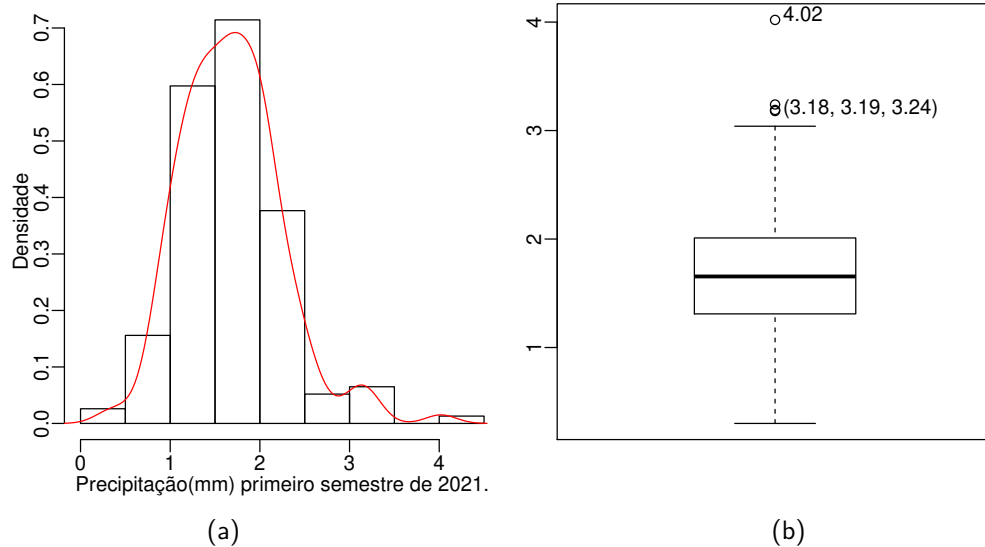
Tabela 18 – Estatísticas descritivas para respostas médias de precipitação entre os meses de Janeiro e Junho de 2021 no estado de Pernambuco, Brasil.

Variável	Min	Máx	Q_1	Q_2	Q_3	mean	Var	Sd	$Cv(\%)$	Ku	AS
Prec(mm)	0,30	4,02	1,31	1,65	2,01	1,69	0,32	0,57	33,58	4,58	0,76

Fonte: Elaborada pelo autor (2022).

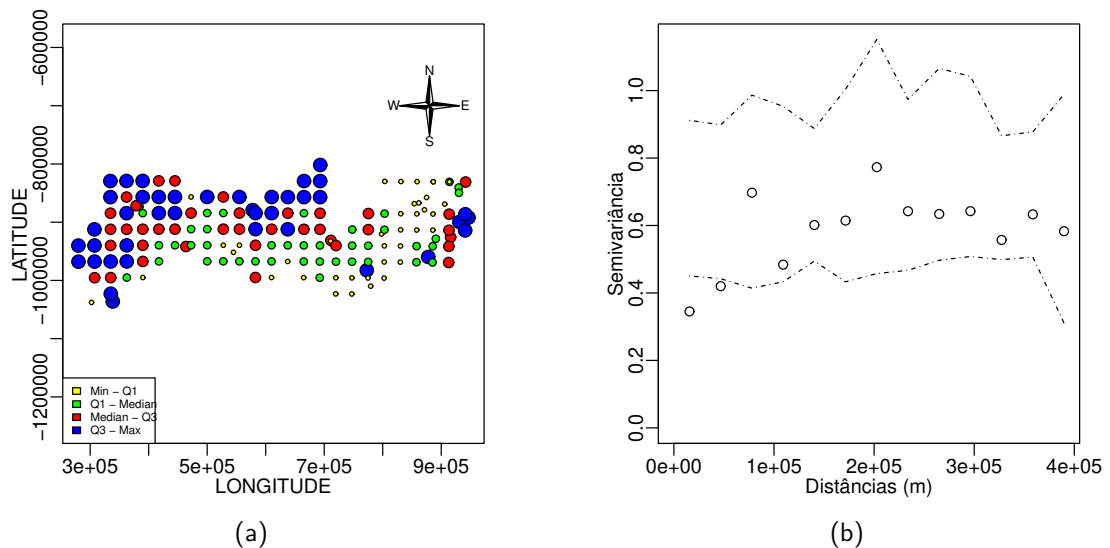
Min: mínimo. Máx: máximo. Q_1 : primeiro quartil. Q_2 : mediana. Q_3 : terceiro quartil. mean: média. Var: variância amostral. Sd: desvio padrão amostral. $Cv(\%)$: coeficiente de variação. Ku: Coeficiente de curtose. AS: Coeficiente de Assimetria. Prec(mm): Precipitação média para primeiro semestre de 2021.

Figura 17 – Histograma de densidade e *BoxPlot* da variável precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil, Figuras 18(a) e 18(b), respectivamente.



Fonte: Elaborada pelo autor (2022).

Figura 18 – *Postplot* e envelope da distribuição espacial da variável precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil, Figuras 19(a) e 19(b), respectivamente.



Fonte: Elaborada pelo autor (2022).

Para modelar os dados médios de precipitação para o primeiro semestre de 2021 no estado de Pernambuco, Brasil, inicialmente, definimos a estrutura de correlação espacial. Para tanto, primeiramente consideramos vetor paramétrico $\varphi = (\phi_1, \phi_2, \phi_3)^T$, com entradas definindo, nesta ordem, o efeito pepita, variância do processo, e escala. Neste estudo, consideramos as

funções de correlação das famílias Wendland, Matérn e modelo Esférico. A estimação do vetor paramétrico é realizada por mínimo quadrados generalizados, com resultados na Tabela 19.

Com isso, e considerando que os dados apresentam uma leve assimetria positiva (0,76), indicamos a distribuição Gama para modelar o comportamento da precipitação. Ainda, da análise dos gráficos 19(a) e 19(b), notamos a presença de aglomerados em algumas regiões. Assim, as coordenadas podem ser utilizadas como algum termo de covariáveis. Ademais, como o objetivo é realizar a modelagem por um modelo de quasi-verossimilhança espacialmente correlacionado, não há necessidade de especificar a forma da distribuição a ser utilizada (no caso a Gama). No entanto, requer que seja indicado a função de variância e uma função de ligação para o modelo de regressão. Sendo assim, é considerada uma função ligação $\eta_i = \frac{1}{\mu_i}$ inversa. Feito isso, desenvolvemos o ajuste de modelos, por meio de regressões polinomiais do primeiro ao quarto grau. Para tanto, consideramos variáveis provenientes da combinação das coordenadas geográficas e altitude, as quais foram, posteriormente, selecionadas a sentimento. Com isso, o modelo de regressão marginal, aqui considerado é dado por:

$$\eta_i = \beta_0 + \beta_1 x(\mathbf{s}_i) + \beta_2 y(\mathbf{s}_j) + \beta_3 x(\mathbf{s}_i)H(\mathbf{s}_i) + \beta_4 x^2(\mathbf{s}_i),$$

em que $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ denotam os parâmetros de regressão, $x(\mathbf{s}_i)$ e $y(\mathbf{s}_j)$, $i, j = 1, \dots, n$, são os pontos de Latitude e Longitude, enquanto H (em metros) denota a altitude da localização $(\mathbf{s}_i, \mathbf{s}_j) \in \mathbb{R}^2$. O produto $x(\mathbf{s}_i)H(\mathbf{s}_i)$ representa a interação entre as variáveis latitude e altitude na localização espacial $(\mathbf{s}_i)$.

Estimados os parâmetros que definem a estrutura de dependência espacial, então, é realizada a modelagem dos dados sob EEG considerando, individualmente, cada uma das estruturas de correlação apresentadas. Para tal, empregamos a função *geeglm* do pacote *geepack* disponível no *Software R Core Team* (2020). Para identificar a estrutura de correlação mais apropriada para os dados, em cada caso, são calculadas as medidas do QIC, CIC, RJC e CNBC. Os valores do RJC indicam para qualquer uma das estruturas de correlação Wendland, para $\kappa = 1, 2, 3$ e Matérn, para $\nu = 0, 5; 1, 5; 2, 5; 3, 5$, por apresentarem o mesmo valor. Neste caso, coincide com o critério CNBC para indicar a estrutura de correlação Wendland com $\kappa = 3$. O critério QIC indica para selecionar a estrutura de correlação do modelo Wendland com $\kappa = 1$. O critério RJC apresentou valores constantes, indicando assim que não é informativo. Por outro lado, os critérios CIC e CNBC indicam modelo Wendland com $\kappa = 3$. Portanto, tendo em vista a boa performance do critério CNBC, tais resultados confirmam a estrutura de correlação da família Wendland com $\kappa = 3$ como apropriada. Os parâmetros para o modelo de

regressão selecionado estão na Tabela 20, com os respectivos erros padrões robustos e *naive* entre parênteses.

Tabela 19 – Estimativas dos parâmetros do modelo geoestatístico para famílias Wendland, Matérn e modelo Esférico, os quais definem a matriz de correlação de trabalho, para variável resposta precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil.

Modelo	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	QIC	CIC	RJC	CNBC
W- $\kappa = 0$	0,1180	0,2216	0,3832	424,2	$4,8 \times 10^{-10}$	1,414	1,236
W- $\kappa = 1$	0,1290	0,1660	0,3773	423,9	$1,1 \times 10^{-9}$	1,414	2,093
W- $\kappa = 2$	0,1179	0,2090	0,3042	437,7	$6,2 \times 10^{-10}$	1,414	1,235
W- $\kappa = 3$	0,1179	0,2205	0,3041	444,6	$2,4 \times 10^{-11}$	1,414	1,107
M- $\nu = 0,5$	0,0220	0,3260	0,0420	427,3	$2,6 \times 10^{-9}$	1,414	1,248
M- $\nu = 1,5$	0,0970	0,2290	0,0290	425,7	$6,1 \times 10^{-10}$	1,414	1,175
M- $\kappa = 2,5$	0,1130	0,2160	0,0240	424,5	$6,2 \times 10^{-10}$	1,414	1,142
M- $\nu = 3,5$	0,1130	0,2110	0,0200	424,3	$6,7 \times 10^{-10}$	1,414	1,121
Esférico	0,0990	0,1800	0,3000	449,8	$2,5 \times 10^{-11}$	1,414	1,410

Fonte: Elaborada pelo autor (2022).

CNBC: Condition-Number-Based Criterion. M - ν : Matérn com suavização κ . W - κ : Wendland com suavização κ .

A partir dos resultados apresentados na Tabela 20, podemos observar que os erros padrões robustos foram maiores que os *naive*. Assumindo 5% de significância, as covariáveis Longitude e interação entre a Latitude e Altitude são marginalmente significativas segundo teste de hipótese Wald. Verificando o gráfico de probabilidade meio-normal com envelope simulado, Figura 20(a), observamos que a grande maioria dos pontos se apresentam dentro do envelope simulado. Desta forma, consideramos que a suposição de uma distribuição marginal Gama para variável resposta precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil é apropriada sob o modelo de quasi-verossimilhança para dados correlacionados da forma que aqui é ajustado.

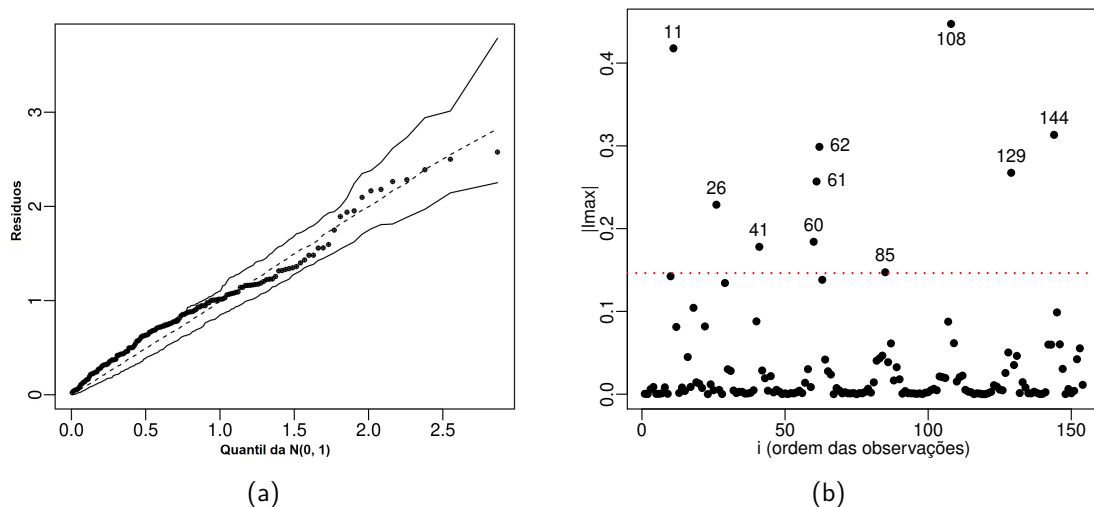
Ademais, no tange as medidas de influência local, segundo um esquema de perturbação na matriz de trabalho, as observações potencialmente influentes para estimativas dos parâmetros foram 11, 26, 41, 60, 61, 62, 85, 108, 129, e 144 correspondente as medidas de precipitação média 2,05; 2,05; 2,35; 1,62; 1,68; 2,32; 2,00; 2,25; 2,12 e 1,78, respectivamente. Para classificar-las como potencialmente influente, consideramos como medida de corte $2 * Sd$ (duas vezes o desvio padrão) das medidas de influência local. Portanto, diante de tais medidas classificadas como potencialmente influente, concluímos que a modelagem dos dados referentes ao Estudo precipitação média deve ser realizado com cautela sob a presença ou não destas observações.

Tabela 20 – Estimativas e erros padrão dos parâmetros do modelo de regressão $\eta_i = \beta_0 + \beta_1 x(s_i) + \beta_2 y(s_j) + \beta_3 x(s_i)H(s_i) + \beta_4 x^2(s_i)$, ajustado com estrutura de correlação Wendland com $\kappa = 3$, para variável resposta precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil.

Parâmetro	Erro padrão		
		Robusto	Naive
$\hat{\beta}_0$	0,891	0,0207	0,0009
$\hat{\beta}_1$	1,72	0,0055	0,0003
$\hat{\beta}_2$	-8,15	0,0360	0,0008
$\hat{\beta}_3$	0,00008	<0,0005	<0,0002
$\hat{\beta}_4$	2,96	< 0,0003	< 0,0007

Fonte: Elaborada pelo autor (2022).

Figura 19 – Gráfico de probabilidade meio-normal com envelope simulado (95% de confiança) do modelo de regressão, com distribuição marginal Gama, ajustado com estrutura de correlação Wendland com $\kappa = 3$, para variável resposta precipitação média para o primeiro semestre de 2021 no estado de Pernambuco, Brasil (20(a)). Gráfico de influência local segundo uma perturbação na matriz de correlação de trabalho (20(b)).



Fonte: Elaborada pelo autor (2022).

3.7 CONSIDERAÇÕES FINAIS DO CAPÍTULO

O que é certo é que selecionar uma estrutura de correlação de trabalho apropriada em GEE tem um papel importante para obter estimativas eficientes dos parâmetros do modelo de regressão marginal. Por isso, não apenas importante, como também vantajoso, propor e compreender critérios que auxiliem na seleção de uma matriz de correlação espacial. Por isso, neste trabalho, um novo critério, nomeado Condition-Number-Based Criterion (CNBC),

foi desenvolvido com base no número de condição das matrizes de variâncias-covariância *naive*, para selecionar uma estrutura de correlação de trabalho espacial apropriada em GEE. Comparamos a nossa proposta com os critérios RJC, QIC e CIC, disponíveis na literatura, por meio de um extenso estudo de simulação. Para gerar as respostas espacialmente correlacionadas foi utilizado o algoritmo *Norta*, com grades regulares de tamanhos 10×10 , 20×20 e uma grade irregular baseada nos conjunto dados reais utilizado como exemplo numérico. Os resultados mostraram que o critério proposto tem melhor desempenho do que o critério RJC, em termos de seleção correta da matriz de correlação, em todos os casos quando considerada respostas espacialmente correlacionada da Poisson e Gama. No caso da distribuição Normal os critérios obtiveram resultados ligeiramente equivalentes. Quando comparado com os critérios QIC e CIC o CNBC obteve melhor desempenho em todos os cenários.

Destacamos que o critério proposto permite que as informações, presentes na matriz de correlação espacial, contribuam para estimativa de todas as partes do modelo de regressão sob EEG, ao contrário dos outros critérios QIC, CIC e RJC. Ademais, neste trabalho focamos na seleção da matriz de correlação de trabalho sob a suposição de que o modelo de média está especificado corretamente. A especificação da estrutura de correlação é baseada em modelos de semivariogramas, utilizando as famílias Wendland, Matérn e modelo Esférico. Assim, investigações futuras do critério CNBC podem incluir a seleção simultânea da matriz de covariância espacial e do preditor linear em EEG.

Finalmente, destacamos que não há métodos atualmente disponíveis na literatura além do proposto neste Capítulo para escolher uma estrutura de correlação espacial adequada em GEE. Os critérios QIC e CIC, revisados neste trabalho, não funcionaram bem quando a verdadeira estrutura de correlação é a família Wendland. Apesar de quando a verdadeira estrutura de correlação é a família Matérn o critério CIC, em alguns casos selecionar a verdadeira estrutura de correlação, concluímos que o mesmo não é adequado para nosso contexto, tendo em vista sua taxa efetiva de seleção inferior a 50%. Por outro lado, o critério QIC funcionou bem nos casos de resposta normal e poisson espacialmente correlacionada com estrutura de correlação da família Matérn. Ainda, apesar do critério RJC ser equivalente ao CNBC para respostas normais, o CNBC foi superior em todos os cenários para resposta Poisson e Gama, tanto para estrutura de correlação da família Wendland quanto da Matérn e modelo Esférico. Portanto, os resultados mostraram que o critério proposto tem desempenho muito superior à critérios popularmente conhecidos na literatura.

4 TÉCNICAS DE DIAGNÓSTICO EM MODELOS QUASI-GAMA AUMENTADA COM ZERO PARA DADOS ESPACIALMENTE CORRELACIONADOS

4.1 RESUMO

De certo, em muitas aplicações, sobretudo no contexto geoestatístico, uma significativa parcela das observações amostradas são zeros, enquanto que a parcela restante assume valores positivos dentro dos números reais. Mas, neste contexto, além dos modelos usuais para variáveis positivas não permitirem acomodar os zeros, também não levam em consideração possíveis dependências entre as observações. Em vista disso, nesta Seção, utilizamos um modelo de regressão de mistura gama aumentada com zeros para analisar dados com dependência espacial e muitos zeros. Para tanto, a construção do modelo é baseada em dois submodelos de componentes ortogonais. A modelagem do primeiro submodelo (para o excesso de zeros) é realizada através de uma regressão logística. Enquanto que para o segundo submodelo, que modela os dados contínuos e positivos, utilizamos um modelo de quasi-verossimilhança da distribuição gama. Incorporamos a dependência espacial, à ambos submodelos, através das equações de estimação generalizadas. Ainda, discutimos os aspectos inferenciais para estimação dos parâmetros de cada submodelo utilizando o algoritmo Expectation-Solution (ES). Além disso, avaliamos o desempenho dos estimadores por um estudo de simulação. Desenvolvemos expressões para os erros-padrão e sugerimos medidas para os resíduos. Adicionalmente, derivamos medidas de influência local baseada na curvatura normal para identificar pontos potencialmente influentes, sob os esquemas de perturbação ponderação de casos, variável resposta da componente contínua e nas variáveis preditoras. Finalmente, ilustramos a metodologia desenvolvida através de um exemplo numérico, no qual foi utilizado dados sobre a quantidade de chuva acumulada (mm) para o mês de agosto do ano de 2021 no estado de Pernambuco, Brasil.

4.2 INTRODUÇÃO

É comum em estatísticas espaciais a análise e/ou modelagem de dados contínuos não negativos. Com isso, são várias as aplicações práticas possíveis. Por exemplo, quando o pesquisador busca entender a dinâmica espacial com o objetivo de compreender o comportamento de fenômenos climáticos, ou prever a ocorrência dos mesmos. Ou ainda, avaliar taxas de in-

cidência epidemiológicas. Particularmente, temos como motivação prática para este estudo, avaliar a distribuição espacial de fenômenos climáticos. Assim, consideramos a quantidade ou ocorrência de chuva em uma dada região geográfica. Portanto, é sabido que a análise pode ser complicada devido a ocorrência de valor zero na amostra, isto é, localizações em que não ocorreu chuva no período em estudo. Desta forma, é permissível imaginar que, dado um vetor aleatório $\mathbf{Z} = (z_1, \dots, z_n)^\top$ contendo n observações amostradas nas localizações espaciais $\mathbf{s}_1, \dots, \mathbf{s}_n \subset \mathbf{S} \in \mathbb{R}^2$, então, $\mathbf{z}_i$, $i = 1, \dots, n$, assume valor zero com probabilidade um se não ocorrer o fenômeno em estudo, e outro valor contínuo e positivo para ocorrência do fenômeno na região em estudo. Em vista disso, é importante o desenvolvimento de modelos que permitam a incorporação dos zeros presentes na variável resposta, à análise dos dados.

Usualmente, os modelos utilizados para modelagem de dados de precipitação pluviométrica (quantidade de chuva), com assimetria positiva são baseados na distribuição Gama. No entanto, modelos com esta característica não compreendem o valor zero, devido o domínio da distribuição Gama. Neste caso, o ajuste com tais modelos não é adequado, pois seria necessário descartar da amostra as observações com valores zeros. Igualmente, modelos para modelagem de dados referente a notificações de casos epidemiológicos não estão disponíveis, sobretudo quando se considera a localização geográfica das amostras, isto é, quando são georreferenciados.

Neste contexto, o objetivo principal desta Seção é propor um modelo que possa acomodar os zeros presentes em amostras de dados com as características acima listadas. Assim, a ideia é desenvolver um modelo de mistura da distribuição degenerada em zero com a distribuição Gama, proposto por Feuerverger (1979), sob dados espacialmente referenciados gerando, assim, um modelo geoestatístico Gama aumentado com zeros. Ou seja, no modelo geoestatístico proposto, a variável resposta é discreta em zero e contínua em $(0, \infty)$.

Os modelos Gama aumentado com zero, definida inicialmente por Feuerverger (1979), foi utilizado para ajustar como modelos de regressão para dados climáticos na Austrália, em que utilizou-se a forma logística para modelar a singularidade distributiva na origem e, a distribuição Gama usual para os preditores da precipitação pluviométrica. A partir de então, diversos trabalhos foram desenvolvidos para modelagem de dados assimétricos inflacionados de zeros. Por exemplo, Heller et al. (2006), Cook, Kieschnick e McCullough (2008) e Venezuela e Artes (2014) propuseram modelos de regressão inflacionados de zeros usando a distribuição Gaussiana Inversa, modelos de regressão beta inflacionados no zero aplicados em análise de decisões de estrutura de capital corporativa, e técnicas de diagnóstico em Equações de Estima-

ção Generalizadas (GEE) com variável resposta inflacionada de zeros, usando as distribuições Gaussiana Inversa e Beta, respectivamente. Os autores () apresentaram um modelo de regressão mista gama aumentada com zero para analisar dados longitudinais com muitos zeros, para o custo médio com sinistros em saúde ocupacional. E, os autores Rodrigues-Motta et al. (2015) estudaram um modelo de efeito misto para respostas positivas aumentadas com zeros.

Sob a perspectiva de dados correlacionados, a teoria dispõe das GEE de Liang e Zeger (1986) e de efeitos aleatórios para explicar a correlação em modelos inflacionados de zeros. Assim, embora existam trabalhos no contexto de dados espaciais, utilizando GEE's para modelar dados com excesso de zeros, por exemplo: Dobbie Melissa J e Welsh (2001) que utilizam as GEE's para modelar dados de contagem inflacionados de zeros; Hall e Zhang (2004) que desenvolveram equações GEE's sob um modelo de mistura para dados longitudinais inflacionados de zeros; Monod (2011) desenvolveram as GEE's para dados de contagem espacial inflacionados a zero; e Bandeira et al. (2015) que desenvolveram as GEE's para modelar dados de contagem espacialmente distribuídos com excesso de zeros e sobredispersão; até o momento da escrita deste trabalho não temos o conhecimento de textos que abordem modelos assimétricos utilizando a distribuição Gama, espacialmente referenciados e com excesso de zeros sob GEE's.

4.3 MODELO QUASI-GAMA ESPACIALMENTE CORRELACIONADO

Seja $Z(s_i) = z_i$ uma realização de um processo estocástico espacial na localização $s_i \in S \subset \mathbb{R}^2$, $i = 1, \dots, n$. Considere ainda a existência de um vetor $\mathbf{x}(s_i) \in \mathbb{R}^p$ de covariáveis. Com isso, aqui estendemos o conceito de distribuições contínuas positiva ajustada no zero para dados espacialmente referenciados. Para tanto, consideramos a função

$$g(z_i; \pi_i, \mu_i, \phi) = \begin{cases} \pi_i, & \text{se } z_i = 0 \\ (1 - \pi_i)f(z_i; \mu_i, \phi), & \text{se } z_i > 0 \end{cases} \quad (4.1)$$

em que $0 < \pi_i < 1$ é a probabilidade de ocorrência do valor zero observado, isto é, $\pi_i = P(z_i = 0)$. A função $f(z_i; \mu_i, \phi)$ é alguma função de probabilidade contínua definida em $(0, \infty)$. Desta forma, (4.1) fica nomeada de função de distribuição ajustada no zero. O interesse aqui é utilizar f como sendo a função densidade da distribuição Gama.

Outrossim, a distribuição Gama ajustada no zero proposta por Feuerverger (1979), é uma extensão da Gama usual permitindo a ocorrência de zeros nos dados. Paralelamente, aqui a

parametrização considerada para distribuição Gama é a proposta por Stasinopoulos, Rigby e Akantziliotou (2008, p.272), ou seja:

$$f(z_i; \mu_i, \phi) = \frac{\exp\left(-\frac{z_i}{\mu_i \phi^2}\right) z_i^{\frac{1}{\phi^2}-1}}{\Gamma\left(\frac{1}{\phi^2}\right) (\mu_i \phi^2)^{\frac{1}{\phi^2}}},$$

com $z_i > 0, \mu_i > 0, \phi > 0$. Sem perda de generalidade, consideramos $\phi = 1$. Neste caso, f pode ser reescrita como:

$$f(z_i; \mu_i, 1) = \frac{\exp\left(-\frac{z_i}{\mu_i}\right)}{\mu_i},$$

logo, temos que para este caso $E(z_i) = \mu_i$ e $Var(z_i) = \mu_i^2$.

Com isso, dado os vetores $\boldsymbol{\pi}, \boldsymbol{\mu}$ de médias para as distribuições binomial e Gama, respectivamente. Então, aqui definimos o modelo Gama aumentado com zeros (ZAGA), com vetor paramétrico $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu})^\top$ por (4.2).

$$g(z_i; \pi_i, \mu_i) = \begin{cases} \pi_i, & \text{se } z_i = 0 \\ (1 - \pi_i) \frac{1}{\mu_i} \exp\left(-\frac{z_i}{\mu_i}\right) & \text{se } z_i > 0, \end{cases} \quad (4.2)$$

em que $E(z_i) = (1 - \pi_i)\mu_i$ e $Var(z_i) = (1 - \pi_i)(\pi_i + 1)\mu_i^2$. Ainda, note que o modelo (4.2) pode ser reescrito em termos da função indicadora, isto é:

$$g(z_i; \pi_i, \mu_i, I_0(z_i)) = \left(\pi_i^{I_0(z_i)} (1 - \pi_i)^{1-I_0(z_i)}\right) \left[\frac{1}{\mu_i} \exp\left(-\frac{z_i}{\mu_i}\right)\right]^{1-I_0(z_i)}, \quad (4.3)$$

em que

$$I_0(z_i) = \begin{cases} 1, & \text{se } z_i = 0 \\ 0, & \text{se } z_i > 0. \end{cases}$$

Neste caso, I_0 é vista como uma variável aleatória binária. Desta forma, o modelo ZAGA pode ser ajustado como um MLG de mistura. Portanto, é permissível assumir que existam funções $h_1(\cdot)$ e $h_2(\cdot)$, estritamente monótonas e duplamente diferenciáveis, tal que:

$$\begin{aligned} h_1(\pi_i) &= \boldsymbol{\zeta} = \mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma} \\ h_2(\mu_i) &= \boldsymbol{\eta} = \mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta}, \end{aligned}$$

em que $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)^\top$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$, são vetores desconhecidos e, $\mathbf{y}(\mathbf{s}_i) = (y_{i1}, \dots, y_{iq})$, $\mathbf{x}(\mathbf{s}_i) = (x_{i1}, \dots, x_{ip})$ são vetores das variáveis preditoras em cada localização espacial $\mathbf{s}_i \in \mathbb{R}^2$, $i = 1, \dots, n$. Ainda, fica definida as matrizes de variáveis preditoras $\mathbf{Y}$ e $\mathbf{X}$ de dimensões $n \times q$ e $n \times p$, respectivamente, com linhas preenchidas pelos vetores $\mathbf{y}(\mathbf{s}_i)$ e $\mathbf{x}(\mathbf{s}_i)$, respectivamente.

Sendo assim, dada uma amostra de n observações independentes amostradas e referenciadas pelas localizações $\mathbf{s}_1, \dots, \mathbf{s}_n$, então, o logaritmo da função verossimilhança para o modelo (4.3) fica dado por:

$$\begin{aligned} \log(g(\mathbf{z}; \boldsymbol{\pi}, \boldsymbol{\mu})) &= \log \left\{ \prod_{i=1}^n \left(\pi_i^{I_0(z_i)} (1 - \pi_i)^{1-I_0(z_i)} \right) \left[\frac{1}{\mu_i} \exp \left(-\frac{z_i}{\mu_i} \right) \right]^{1-I_0(z_i)} \right\} \\ &= l_1(\boldsymbol{\gamma}) + l_2(\boldsymbol{\beta}), \end{aligned}$$

em que

$$\begin{aligned} l_1(\boldsymbol{\gamma}) &= \sum_{i=1}^n \{ I_0(z_i) \log(\pi_i) + (1 - I_0(z_i)) \log(1 - \pi_i) \}, \\ l_2(\boldsymbol{\beta}) &= \sum_{i=1}^n (1 - I_0(z_i)) \left\{ -\frac{z_i}{\mu_i} - \log(\mu_i) \right\}, \end{aligned}$$

com $\pi_i = h_1^{-1}(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})$ e $\mu_i = h_2^{-1}(\mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta})$. Aqui, consideramos a função de ligação ($h_1(\cdot)$) logística para o parâmetro de excesso de zeros, isto é:

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma},$$

e para os parâmetros de regressão consideramos $\log(\mu_i) = \mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta}$.

Com isso, nota-se que $l_1(\boldsymbol{\gamma})$ e $l_2(\boldsymbol{\beta})$ dependem unicamente de $\boldsymbol{\gamma}$ e $\boldsymbol{\beta}$, respectivamente, portanto, são separáveis. Sendo assim, a ideia aqui é substituir o logaritmo da verossimilhança para o modelo Gama ($l_2(\boldsymbol{\beta})$) pela sua Quasi-verossimilhança. Logo, estabelecemos um modelo Quasi-Gama aumentada com zero dado por (4.4).

$$Q(\mathbf{z}; \boldsymbol{\pi}, \boldsymbol{\mu}) = \sum_{i=1}^n \{ z_i^* \log(\pi_i) + (1 - z_i^*) \log(1 - \pi_i) \} + \sum_{i=1}^n (1 - z_i^*) Q^*(z_i; \mu_i), \quad (4.4)$$

em que $z_i^* = I_0(z_i)$, enquanto que $Q_1(z_i; \mu_i)$ é a Quasi-verossimilhança da distribuição Gama com parâmetro de dispersão $\phi = 1$, e dada por:

$$Q^*(z_i; \mu_i) = \int_{z_i}^{\mu_i} \frac{z_i - t}{t^2} dt = 1 + \log(z_i) - \left(\frac{z_i}{\mu_i} + \log(\mu_i) \right).$$

Desta maneira, e a partir da função de ligação para o excesso de zeros, temos que $\pi_i = \frac{\exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})}{1 + \exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})}$ e $1 - \pi_i = \frac{1}{1 + \exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})}$. Logo, substituindo em (4.4), obtemos

$$\begin{aligned} Q(\boldsymbol{\theta}|\mathbf{z}, \mathbf{z}^*) &= \sum_{i=1}^n \left\{ z_i^* \log \left(\frac{\exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})}{1 + \exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})} \right) + (1 - z_i^*) \log \left(\frac{1}{1 + \exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})} \right) \right. \\ &\quad \left. + (1 - z_i^*) \left[1 + \log(z_i) - \left(\frac{z_i}{\mu_i} + \log(\mu_i) \right) \right] \right\} \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n \{z_i^* \mathbf{y}(\mathbf{s}_i) \boldsymbol{\gamma} - \log(1 + \exp(\mathbf{y}(\mathbf{s}_i) \boldsymbol{\gamma})) \\
&+ (1 - z_i^*) \left[1 + \log(z_i) - \left(\frac{z_i}{\mu_i} + \log(\mu_i) \right) \right] \}, \tag{4.5}
\end{aligned}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$. Diante disso, e utilizando a esperança condicional, temos que

$$\begin{aligned}
E[Q(\boldsymbol{\theta} | \mathbf{z}, \mathbf{z}^*) | \mathbf{z}] &= \sum_{i=1}^n \{ \hat{z}_i^* \mathbf{y}(\mathbf{s}_i) \boldsymbol{\gamma} - \log(1 + \exp(\mathbf{y}(\mathbf{s}_i) \boldsymbol{\gamma})) \\
&+ (1 - \hat{z}_i^*) \left[1 + \log(z_i) - \left(\frac{z_i}{\mu_i} + \log(\mu_i) \right) \right] \},
\end{aligned}$$

em que $\hat{z}_i^*$ é obtido aplicando o teorema de Bayes, ou seja

$$\begin{aligned}
\hat{z}_i^* &= E[z_i^* | z_i, \boldsymbol{\gamma}, \boldsymbol{\beta}] \\
&= \frac{P(z_i^*) P(z_i | z_i^*)}{P(z_i)} \\
&= \frac{\pi_i^{z_i^*} (1 - \pi_i)^{1-z_i^*} g(z_i; \pi_i, \mu_i, z_i^*)^{1-z_i^*}}{\left[\pi_i + (1 - \pi_i) \frac{1}{\mu_i} \exp\left(-\frac{z_i}{\mu_i}\right) \right]}.
\end{aligned}$$

Portanto, tendo em vista que z_i^* é uma realização de uma variável aleatória a partir da função indicadora, e utilizando a função e ligação especificada para o excesso de zeros, com algumas operações algébrica temos

$$\hat{z}_i^* = \begin{cases} \left[1 + \exp(-\mathbf{y}(\mathbf{s}_i) \boldsymbol{\gamma}) \mu_i^{-1} \right]^{-1}, & \text{se } z_i = 0, \\ 0, & \text{se } z_i > 0. \end{cases} \tag{4.6}$$

Sendo assim, aplicando a técnica de Liang e Zeger (1986), as equações de estimação sob independência para modelar a média e o excesso de zeros são dadas por:

$$\mathbf{U}_{\boldsymbol{\gamma}} = \frac{\partial Q(\mathbf{z}; \boldsymbol{\pi}, \boldsymbol{\mu})}{\partial \boldsymbol{\gamma}} = \mathbf{D}_1^\top \left(\mathbf{A}_1^{1/2} \mathbf{I} \mathbf{A}_1^{1/2} \right)^{-1} (\mathbf{z}^* - \boldsymbol{\pi}) \tag{4.7}$$

$$\mathbf{U}_{\boldsymbol{\beta}} = \frac{\partial Q(\mathbf{z}; \boldsymbol{\pi}, \boldsymbol{\mu})}{\partial \boldsymbol{\beta}} = \mathbf{D}_2^\top \left(\mathbf{A}_2^{1/2} \mathbf{I} \mathbf{A}_2^{1/2} \right)^{-1} \mathbf{Z}^* (\mathbf{z} - \boldsymbol{\mu}), \tag{4.8}$$

em que $\mathbf{D}_1$ e $\mathbf{D}_2$ são matrizes de ordem $n \times p$ com elementos dados por $\frac{\partial \pi_i}{\partial \gamma_j}$ e $\frac{\partial \mu_i}{\partial \beta_j}$, respectivamente, com $i = 1, \dots, n$ e $j = 1, \dots, p$; $\mathbf{A}_1 = \text{diag}(V(\pi_1), \dots, V(\pi_n))$ e $\mathbf{A}_2 = \text{diag}(V(\mu_1), \dots, V(\mu_n))$; $\mathbf{I}$ é a matriz identidade de ordem n ; $\mathbf{z}^* = (z_1^*, \dots, z_n^*)$ é um vetor de realizações da variável aleatória indicadora; $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$ o vetor de probabilidades de ocorrência do valor zero observado; e $\mathbf{Z}^* = \text{diag}(1 - z_1^*, \dots, 1 - z_n^*)$.

Para acomodar a dependência espacial, aqui, substituímos a matriz identidade nas Equações (4.7) e (4.8), por uma matriz de correlação espacial $\mathbf{R}(\phi)$ obtida por alguma modelo de

semivariograma. O vetor $\phi = (\phi_1, \phi_2, \phi_3)^\top$ caracteriza a estrutura de dependência espacial. Com isso, as funções escores são reescritas como em (4.9) e (4.10).

$$\mathbf{U}_\gamma = \mathbf{Y}^\top \mathbf{K}_1 (\mathbf{A}_1^{1/2} \mathbf{R} \mathbf{A}_1^{1/2})^{-1} (\mathbf{z}^* - \boldsymbol{\pi}) \quad (4.9)$$

$$\mathbf{U}_\beta = \mathbf{X}^\top \mathbf{K}_2 (\mathbf{A}_2^{1/2} \mathbf{R} \mathbf{A}_2^{1/2})^{-1} \mathbf{Z}^* (\mathbf{z} - \boldsymbol{\mu}), \quad (4.10)$$

em que $\mathbf{K}_1 = \text{diag}(\pi_1(1 - \pi_1), \dots, \pi_n(1 - \pi_n))$, $\mathbf{K}_2 = \text{diag}(\exp(\mathbf{x}(\mathbf{s}_1)\boldsymbol{\beta}), \dots, \exp(\mathbf{x}(\mathbf{s}_n)\boldsymbol{\beta}))$, $\mathbf{A}_1 = \text{diag}(\pi_1(1 - \pi_1), \dots, \pi_n(1 - \pi_n))$ e $\mathbf{A}_2 = \text{diag}(\exp(\mathbf{x}(\mathbf{s}_1)\boldsymbol{\beta})^2, \dots, \exp(\mathbf{x}(\mathbf{s}_n)\boldsymbol{\beta})^2)$. Note que,

$$\pi_i(1 - \pi_i) = \frac{\exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma})}{(1 + \exp(\mathbf{y}(\mathbf{s}_i)\boldsymbol{\gamma}))^2}, \quad i = 1, \dots, n.$$

Por conseguinte, a função de estimação fica dada por (4.11).

$$\mathbf{U}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{U}_\gamma \\ \mathbf{U}_\beta \end{bmatrix}. \quad (4.11)$$

4.3.1 Estimação dos parâmetros

Para estimação dos vetores $\boldsymbol{\gamma}$ e $\boldsymbol{\beta}$ nos modelos Quasi-gama aumentada com zero para dados espacialmente referenciados, devemos considerar a função escore (4.11). Consideramos ainda que $\phi = (\phi_1, \phi_2, \phi_3)^\top$ é fixado e conhecido.

Em geral, a estimação dos parâmetros em modelos GLM's de mistura é realizada pelos algoritmo EM de Dempster, Laird e Rubin (1977). Contudo, sua aplicação exige o uso da função verossimilhança, portanto, não a podemos utilizar em nosso caso. Com isso, utilizaremos uma generalização do algoritmo EM proposta por Rosen, Jiang e Tanner (2000). Esta técnica, a autora a definiu como *Expectation-Solution* (ES) e, utilizou para explicar a correlação no contexto de medidas repetidas em EEG. Ainda, a autora advoga que se o algoritmo convergir, então, convergirá para um estimador não viesado, consistente e assintoticamente normal. Sua aplicação é de forma similar ao algoritmo EM, substituindo o passo M por um passo S que requer a solução de um sistema de equações generalizadas.

Como visto anteriormente, consideramos inicialmente a realização de uma variável aleatória

$$Z^*(\mathbf{s}_i) = \begin{cases} 1, & \text{se } z_i = 0 \\ 0, & \text{se } z_i \sim \text{Gama}(\cdot; \mu_i, \phi = 1). \end{cases}$$

Sendo assim, a realização $Z^*(s_i) = z_i^*$ da variável aleatória Z^* caracteriza os dados não observados. Então, o algoritmo ES é utilizado para maximizar o modelo Quasi-Gama aumentado com zero dado em (4.4), alternando entre o passo E , em que os dados não observados são estimados a partir da esperança condicional e $\theta = (\gamma^\top, \beta^\top)^\top$ e o passo S resolvendo os sistemas de equações (4.9) e (4.10) com $\mathbf{z}^*$ fixo. A solução dos sistemas (4.9) e (4.10) é obtida a partir de uma modificação no algoritmo Fisher scoring. Com isso, temos que

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} - \left\{ E_\beta \left[\frac{\partial \mathbf{U}_\beta}{\partial \beta^\top} \right] \right\}^{-1} \mathbf{U}_\beta \\ &= \beta^{(k)} - \left\{ E_\beta \left[-\mathbf{D}_2^\top \mathbf{V}^{-1} \mathbf{Z}^* \frac{\partial \boldsymbol{\mu}}{\partial \beta^\top} \right] \right\}^{-1} \mathbf{U}_\beta \\ &= \beta^{(k)} + \left((\mathbf{D}_2^{(k)})^\top (\mathbf{V}^{(k)})^{-1} \mathbf{Z}^* \mathbf{D}_2^{(k)} \right)^{-1} (\mathbf{D}_2^{(k)})^\top (\mathbf{V}^{(k)})^{-1} \mathbf{Z}^* (\mathbf{z} - \boldsymbol{\mu}^{(k)}),\end{aligned}\quad (4.12)$$

em que $\mathbf{V} = \mathbf{A}_2^{1/2} \mathbf{R} \mathbf{A}_2^{1/2}$. Similarmente, os parâmetros de regressão para o excesso de zeros são obtidos por:

$$\gamma^{(k+1)} = \gamma^{(k)} + \left((\mathbf{D}_1^{(k)})^\top (\mathbf{V}_1^{(k)})^{-1} \mathbf{D}_1^{(k)} \right)^{-1} (\mathbf{D}_1^{(k)})^\top (\mathbf{V}_1^{(k)})^{-1} (\mathbf{z}^* - \boldsymbol{\pi}^{(k)}), \quad (4.13)$$

neste caso $\mathbf{V}_1 = \mathbf{A}_1^{1/2} \mathbf{R} \mathbf{A}_1^{1/2}$. Feito isto, as estimativas para β e γ são obtidas utilizando o algoritmo (2).

Algorithm 2 Passos para aplicar o algoritmo ES

Passo E - Defina o vetor de parâmetros iniciais $(\gamma^{(0)}, \beta^{(0)})$. Então, iniciamos o processo para estimar $z_i^{*(k)}$ da k -ésima iteração a partir da esperança condicional $\hat{z}_i^{*(k)} = E[z_i^* | z_i, \gamma^{(k)}, \beta^{(k)}]$, com $k = 0, 1, \dots$. Assim, a partir de (4.6), z_i^* no k -ésimo passo de iteração é obtido por

$$\hat{z}_i^{*(k)} = \begin{cases} \left[1 + \exp(-\mathbf{y}(s_i) \gamma) \mu_i^{-1} \right]^{-1}, & \text{se } z_i = 0, \\ 0, & \text{se } z_i > 0. \end{cases}$$

Passo S para β - Defina $(\mathbf{V}_*^{(k)})^{-1} = (\mathbf{A}_2^{1/2} \mathbf{R}(\phi) \mathbf{A}_2^{1/2})^{-1} (\mathbf{Z}^*)^{(k)}$. Então, estimamos o vetor paramétrico (β) para o modelo de regressão, a partir do processo iterativo (4.12).

Passo S para γ - Defina $(\mathbf{V}_1^{(k)})^{-1} = (\mathbf{A}_1^{1/2} \mathbf{R}(\phi) \mathbf{A}_1^{1/2})^{-1}$. Então, estimamos (γ) pelo processo iterativo (4.13).

4.3.2 Erros padrão

Os autores Liang e Zeger (1986) determinam expressões para os estimadores *naïve* e robusto para os parâmetros de regressão sob GEE. Então, para o modelo Quasi-gama aumentada

com zero e espacialmente correlacionados sob GEE, os erros-padrão são obtidos de forma similar, considerando que na convergência do processo temos que:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^\top \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{W}} \hat{\mathbf{T}}, \\ \hat{\gamma} &= (\mathbf{Y}^\top \widehat{\mathbf{W}}_1 \mathbf{Y})^{-1} \mathbf{Y}^\top \widehat{\mathbf{W}}_1 \hat{\mathbf{T}}_1,\end{aligned}$$

em que $\widehat{\mathbf{W}} = \widehat{\mathbf{K}}_2 \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{Z}}^* \widehat{\mathbf{K}}_2$, $\widehat{\mathbf{W}}_1 = \widehat{\mathbf{K}}_1 \widehat{\mathbf{V}}_1^{-1} \widehat{\mathbf{K}}_1$, $\hat{\mathbf{T}} = \hat{\mathbf{X}} \hat{\beta} + \widehat{\mathbf{K}}_2^{-1} (\mathbf{z} - \hat{\boldsymbol{\mu}})$ e $\hat{\mathbf{T}}_1 = \mathbf{Y} \hat{\gamma} + \widehat{\mathbf{K}}_1^{-1} (\mathbf{z}^* - \hat{\boldsymbol{\pi}})$. Dessa forma, assumindo $\hat{\boldsymbol{\mu}}$ e $\hat{\boldsymbol{\pi}}$ aproximadamente os verdadeiros vetores de médias para o componente contínuo e discreto, respectivamente, então, os estimadores para as variâncias de β e γ são obtidas por:

$$\begin{aligned}\widehat{Var}(\hat{\beta}) &= (\mathbf{X}^\top \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{W}} \widehat{\mathbf{K}}_2^{-1} Var(\mathbf{z}) \widehat{\mathbf{K}}_2^{-1} \widehat{\mathbf{W}} \mathbf{X} (\mathbf{X}^\top \widehat{\mathbf{W}} \mathbf{X})^{-1} \\ \widehat{Var}(\hat{\gamma}) &= (\mathbf{Y}^\top \widehat{\mathbf{W}}_1 \mathbf{Y})^{-1} \mathbf{Y}^\top \widehat{\mathbf{W}}_1 \widehat{\mathbf{K}}_1^{-1} Var(\mathbf{z}^*) \widehat{\mathbf{K}}_1^{-1} \widehat{\mathbf{W}}_1 \mathbf{Y} (\mathbf{Y}^\top \widehat{\mathbf{W}}_1 \mathbf{Y})^{-1}.\end{aligned}$$

Nesta conjuntura, assumindo $\mathbf{R}(\phi)$ como sendo a verdadeira matriz de correlação para $\mathbf{z}$, então, portanto, $Var(\mathbf{z}) = \mathbf{A}_2^{1/2} \mathbf{R}(\phi) \mathbf{A}_2^{1/2}$ e $Var(\mathbf{z}^*) = \mathbf{A}_1^{1/2} \mathbf{R}(\phi) \mathbf{A}_1^{1/2}$. Com isso, a estimativa para $\widehat{Var}(\hat{\beta})$ pode ser obtida por:

$$\begin{aligned}\widehat{Var}(\hat{\beta}) &= (\mathbf{X}^\top \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{W}} \widehat{\mathbf{K}}_2^{-1} \widehat{Var}(\mathbf{z})_N \widehat{\mathbf{K}}_2^{-1} \widehat{\mathbf{W}} \mathbf{X} (\mathbf{X}^\top \widehat{\mathbf{W}} \mathbf{X})^{-1} \\ \widehat{Var}(\hat{\gamma}) &= (\mathbf{Y}^\top \widehat{\mathbf{W}}_1 \mathbf{Y})^{-1} \mathbf{Y}^\top \widehat{\mathbf{W}}_1 \widehat{\mathbf{K}}_1^{-1} \widehat{Var}(\mathbf{z}^*)_N \widehat{\mathbf{K}}_1^{-1} \widehat{\mathbf{W}}_1 \mathbf{Y} (\mathbf{Y}^\top \widehat{\mathbf{W}}_1 \mathbf{Y})^{-1},\end{aligned}$$

em que $\widehat{Var}(\mathbf{z})_N = \widehat{\mathbf{A}}_2^{1/2} \mathbf{R}(\phi) \widehat{\mathbf{A}}_2^{1/2}$ e $\widehat{Var}(\mathbf{z}^*)_N = \widehat{\mathbf{A}}_1^{1/2} \mathbf{R}(\phi) \widehat{\mathbf{A}}_1^{1/2}$ são estimativas para variâncias do tipo *naive*. Contudo, na prática a matriz $\mathbf{R}(\phi)$ é desconhecida, portanto, neste caso, são obtidas estimativas de variâncias robustas, substituindo $\widehat{Var}(\mathbf{z})_N$ por $\widehat{Var}(\mathbf{z})_{rob} = (\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^\top$ e $\widehat{Var}(\mathbf{z}^*)_{rob} = (\mathbf{z}^* - \boldsymbol{\pi})(\mathbf{z}^* - \boldsymbol{\pi})^\top$, para os componentes da parte contínua e discreta (excesso de zeros), respectivamente. Então, mediante o exposto, podemos obter as estimativas para os erros padrão *naive* e robusto, para ambos os componentes do modelo.

4.4 TÉCNICAS DE DIAGNÓSTICO

O objetivo da análise de diagnóstico é verificar, a partir do modelo ajustado, possíveis afastamentos das pressuposições do modelo. Desta forma, é essencial o estudo de algum tipo de resíduo para auxiliar na verificação das suposições assumidas. A análise dos resíduos constitui uma importante ferramenta para avaliar e/ou identificar o impacto da presença de pontos discrepantes nas inferências sobre o modelo. Diante disso, aqui estendemos os conceitos de

resíduos e padronizados para o modelo Quasi-gama aumentado com zero e espacialmente correlacionado. Com isso, é possível avaliar a precisão dos resultados inferenciais a respeito do modelo proposto, ou seja, examinar a sensibilidade do método utilizado a pequenas violações das pressuposições. Adicionalmente, desenvolvemos medidas de influência local, sob um esquema de perturbação de ponderação de casos, para o modelo Quasi-gama aumentado com zero e espacialmente correlacionada.

4.4.1 Resíduos

Sendo assim, e considerando a ortogonalidade dos parâmetros dos modelos para o componente discreto e contínuo, é permissível que os resíduos sejam obtido separados. Posto que a modelagem do excesso de zeros é realizada por ajuste de uma resposta binária, então, neste caso, utilizamos os resíduos de Pearson para resposta binária espacialmente correlacionada de forma similar a definida em (ALBERT; MCSHANE, 1995), ou seja:

$$r_i^{(1)} = \frac{I_0(z_i) - \pi_i}{\sqrt{\pi_i(1 - \pi_i)}},$$

em que $I_0(z_i)$ denota a função indicadora.

Para os resíduos do componente contínuo, temos que levar em consideração, também, as estimativas para a probabilidade de uma observação assumir valor diferente de zero. Desta maneira, a proposta aqui é realizar uma extensão dos resíduos definidos para modelos beta inflacionados, proposto por Martinez (2008), Ospina e Ferrari (2012). Portanto, dada a medida $(1 - \pi_i)$, a probabilidade da observação z_i amostrada na localização s_i , $i = 1, \dots, n$ assumir um valor diferente de zero, então definimos o resíduo para z_i como sendo:

$$\begin{aligned} r_i^{(2)} &= \frac{r_i}{\sqrt{1 - \hat{\pi}_i}} \\ &= \frac{z_i - \mu_i}{\sqrt{\hat{V}_{ii}(1 - \hat{\pi}_i)}}, \end{aligned}$$

em que $\hat{V}_{ii}$ é o i -ésimo elemento da diagona principal da matriz $\mathbf{V}$.

Portanto, definimos um resíduo geral para o modelo Quasi-gama aumentada com zero e espacialmente correlacionado, como sendo:

$$r_i = \begin{cases} r_i^{(1)}, & se \quad z_i = 0 \\ r_i^{(2)}, & se \quad z_i > 0, \quad i = 1, \dots, n. \end{cases}$$

4.4.2 Medidas de Influência local

O objetivo desta subseção, é desenvolver medidas de diagnóstico de influência local tal que seja possível investigar se a presença ou não de observações potencialmente influentes, sob pequenas perturbações impostas ao modelo, exercem mudanças significativas nos resultados inferenciais. Com isso, a ideia é estender a metodologia proposta por Cook (1986) a partir do conceito de influência local generalizada apresentada por Venezuela, Botter e Sandoval (2007), para derivar medidas de influência local para o modelo aqui proposto. Para tanto, assumimos fixa a matriz de correlação espacial $\mathbf{R}(\phi)$. Então, seja $\mathcal{F}(\boldsymbol{\theta})$ uma função de ajuste duplamente diferenciável, tal que

$$\mathbf{U}(\boldsymbol{\theta}) = \left[\frac{\partial \mathcal{F}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0},$$

em que $\mathbf{U}(\boldsymbol{\theta})$ é a função de estimação definida em (4.11) para o modelo Quasi-gama aumentado com zero e espacialmente correlacionado. Assim sendo, denotamos a função de afastamento em relação ao modelo perturbado por:

$$FD_{\boldsymbol{\omega}} = 2 \left\{ \mathcal{F}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) - \mathcal{F}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}) \right\},$$

em $\boldsymbol{\omega} = (\omega_1, \dots, \omega_l)^\top$ representa um vetor de perturbação, tal que $\boldsymbol{\omega} \subset \boldsymbol{\Omega} \in \mathbb{R}^l$, sendo $\boldsymbol{\Omega}$ o conjunto de perturbações. Ainda, os vetores $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ são as estimativas de $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$ sob os modelos postulado e perturbado, respectivamente.

Desta forma, a proposta consiste em estudar o comportamento de $FD_{\boldsymbol{\omega}}$ localmente, para qualquer vetor $\boldsymbol{\omega}$ na vizinhança de algum vetor de não perturbação $\boldsymbol{\omega}_0$ ($FD_{\boldsymbol{\omega}_0} = 0$). À vista disso, tal qual sugere Cook (1986), devemos investigar a curvatura normal referente ao gráfico de $FD_{\boldsymbol{\omega}_{0+a\mathbf{d}}}$ versus a , com $a \in \mathbb{R}$ e $\mathbf{d}$ um vetor unitário ($\|\mathbf{d}\| = 1$). A curvatura normal é obtida por:

$$C_{\mathbf{d}} = -2 \left(\mathbf{d}^\top \boldsymbol{\Delta}^\top \ddot{\mathcal{F}}^{-1} \boldsymbol{\Delta} \mathbf{d} \right),$$

em que, como estamos considerando $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$, com $\boldsymbol{\gamma}$ e $\boldsymbol{\beta}$ de dimensões $p \times 1$, então, $\boldsymbol{\Delta}$ é uma matriz de dimensão $2p \times n$ com elementos dados por $\frac{\partial^2 \mathcal{F}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \boldsymbol{\omega})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^\top}$. Com isso, identificamos observações que poderão exercer influência considerável na função de afastamento $FD_{\boldsymbol{\omega}}$, a partir da máxima direção $\mathbf{d}_{max}$ da curvatura normal, obtida pelo autovalor correspondente ao maior autovetor da matriz

$$\mathbf{B} = -\boldsymbol{\Delta}^\top \ddot{\mathcal{F}}^{-1} \boldsymbol{\Delta},$$

em que $\ddot{\mathcal{F}} = \frac{\partial^2 \mathcal{F}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$ é uma matriz bloco diagonal de dimensão $2p \times 2p$ avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Além disso, dada a separabilidade de (4.11), podemos escrever Δ como:

$$\Delta = \begin{pmatrix} \Delta_\gamma = \frac{\partial^2 \mathcal{F}(\gamma|\hat{\gamma}, \omega)}{\partial \gamma \partial \omega^\top} \Big|_{\gamma=\hat{\gamma}, \omega=\omega_0} \\ \Delta_\beta = \frac{\partial^2 \mathcal{F}(\beta|\hat{\beta}, \omega)}{\partial \beta \partial \omega^\top} \Big|_{\beta=\hat{\beta}, \omega=\omega_0} \end{pmatrix}.$$

Note que Δ depende do tipo de perturbação a ser considerada. Ainda, temos que:

$$\ddot{\mathcal{F}} = \frac{\partial^2 \mathcal{F}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \ddot{\mathcal{F}}_{\gamma\gamma} = \frac{\partial^2 \mathcal{F}(\gamma|\hat{\gamma})}{\partial \gamma \partial \gamma^\top} & \mathbf{0} \\ \mathbf{0} & \ddot{\mathcal{F}}_{\beta\beta} = \frac{\partial^2 \mathcal{F}(\beta|\hat{\beta})}{\partial \beta \partial \beta^\top} \end{pmatrix},$$

Destarte, para o modelo Quasi-gama aumentada com zero e espacialmente correlacionado, temos que:

$$\begin{aligned} \ddot{\mathcal{F}}_{\gamma\gamma} &= \frac{\partial}{\partial \gamma} \left\{ \frac{\partial \mathcal{F}(\gamma|\hat{\gamma})}{\partial \gamma^\top} \right\} \\ &= \mathbf{Y}^\top \left\{ \frac{\partial (\mathbf{K}_1 \mathbf{V}_1^{-1})}{\partial \gamma_j} (\mathbf{z}^* - \boldsymbol{\pi}) - \mathbf{K}_1 \mathbf{V}_1^{-1} \frac{\partial \boldsymbol{\pi}}{\partial \gamma_j} \right\} \quad j = 1, \dots, p, \end{aligned}$$

em que $\mathbf{V}_1 = \mathbf{A}_1^{1/2} \mathbf{R}(\phi) \mathbf{A}_1^{1/2}$ e $\frac{\partial \boldsymbol{\pi}}{\partial \gamma_j} = \mathbf{K}_1 \mathbf{y}_j$, com $\mathbf{y}_j$ denotando a j -ésima coluna da matriz de preditores do excesso de zeros $\mathbf{Y}$. Outrossim, $\frac{\partial (\mathbf{K}_1 \mathbf{V}_1^{-1})}{\partial \gamma_j} = \frac{\partial \mathbf{K}_1}{\partial \gamma_j} \mathbf{V}_1^{-1} + \mathbf{K}_1 \frac{\partial \mathbf{V}_1^{-1}}{\partial \gamma_j}$, com $\frac{\partial \mathbf{V}_1^{-1}}{\partial \gamma_j} = \frac{\partial \mathbf{V}_1}{\partial \gamma_j} \mathbf{V}_1^{-1} \frac{\partial \mathbf{V}_1}{\partial \gamma_j}$ e $\frac{\partial \mathbf{V}_1}{\partial \gamma_j} = \frac{\partial \mathbf{A}_1^{1/2}}{\partial \gamma_j} \mathbf{R}(\phi) \mathbf{A}_1^{1/2} + \mathbf{A}_1^{1/2} \mathbf{R}(\phi) \frac{\partial \mathbf{A}_1^{1/2}}{\partial \gamma_j}$. Ainda,

$$\begin{aligned} \frac{\partial \mathbf{K}_1}{\partial \gamma_j} &= \text{diag} \left(\frac{\exp(\zeta_i) y_{ij} (1 + \exp(\zeta_i) \gamma)^2 - 2 \exp(\zeta_i) (1 + \exp(\zeta_i)) \exp(\zeta_i) y_{ij}}{(1 + \exp(\zeta_i))^4} \right) \\ &= \text{diag} \left(\exp(\zeta_i) \frac{(1 - \exp(\zeta_i)) y_{ij}}{(1 + \exp(\zeta_i))^3} \right) \\ &= \text{diag} (\pi_i (1 - \pi_i) (1 - 2\pi_i) y_{ij}), \quad i = 1, \dots, n, \quad j = 1, \dots, p, \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathbf{A}_1^{1/2}}{\partial \gamma_j} &= \frac{\partial}{\partial \gamma_j} \left\{ \text{diag} ((\pi_i (1 - \pi_i))^{1/2}) \right\} \\ &= \text{diag} \left(\frac{1}{2} (\pi_i (1 - \pi_i))^{-1/2} \pi_i (1 - \pi_i) (1 - 2\pi_i) y_{ij} \right) \\ &= \text{diag} \left(\frac{1}{2} (\pi_i (1 - \pi_i))^{1/2} (1 - 2\pi_i) y_{ij} \right). \end{aligned}$$

Analogamente, os elemento da matriz $\ddot{\mathcal{F}}_{\beta\beta}$ são obtidos por:

$$\ddot{\mathcal{F}}_{\beta\beta} = \frac{\partial}{\partial \beta} \left\{ \frac{\partial \mathcal{F}(\beta|\hat{\beta})}{\partial \beta^\top} \right\}$$

$$= \mathbf{X}^\top \left\{ \frac{\partial(\mathbf{K}_2 \mathbf{V}_2^{-1})}{\partial \beta_j} \mathbf{Z}^* (\mathbf{z} - \boldsymbol{\mu}) - \mathbf{K}_2 \mathbf{V}_2^{-1} \mathbf{Z}^* \frac{\partial \boldsymbol{\mu}}{\partial \beta_j} \right\} \quad j = 1, \dots, p,$$

em que $\mathbf{V}_2 = \mathbf{A}_2^{1/2} \mathbf{R}(\phi) \mathbf{A}_2^{1/2}$ e $\frac{\partial \boldsymbol{\mu}}{\partial \beta_j} = \mathbf{K}_2 \mathbf{x}_j$, com $\mathbf{x}_j$ denotando a j -ésima coluna da matriz de preditores $\mathbf{X}$. Ainda,

$$\frac{\partial \mathbf{K}_2}{\partial \beta_j} = \text{diag}(\exp(\eta_i) x_{ij}) = \frac{\partial \mathbf{A}_2^{1/2}}{\partial \beta_j}.$$

Desta forma, a seguir apresentamos as expressões para a matriz Δ já avaliadas em $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, segundo cada um dos cenários de perturbação considerados.

4.4.2.1 Esquema de perturbação ponderação de casos

Seja um vetor de perturbação $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^\top$ e, $\boldsymbol{\omega}_0 = (1, \dots, 1)^\top$ de ordem $n \times 1$ um vetor de não perturbação. Com isso, as equações de estimativa para o modelo quasi-gama aumentado com zero e espacialmente correlacionado, perturbado, são dadas por:

$$\begin{aligned} \mathbf{U}(\boldsymbol{\gamma}|\boldsymbol{\omega}) &= \mathbf{Y}^\top \mathbf{K}_1 \left(\mathbf{A}_1^{1/2} \mathbf{R}(\phi) \mathbf{A}_1^{1/2} \right)^{-1} \text{diag}(\boldsymbol{\omega}) (\mathbf{z}^* - \boldsymbol{\pi}) \\ \mathbf{U}(\boldsymbol{\beta}|\boldsymbol{\omega}) &= \mathbf{X}^\top \mathbf{K}_2 \left(\mathbf{A}_2^{1/2} \mathbf{R}(\phi) \mathbf{A}_2^{1/2} \right)^{-1} \mathbf{Z}^* \text{diag}(\boldsymbol{\omega}) (\mathbf{z} - \boldsymbol{\mu}). \end{aligned}$$

Desta forma, é fácil ver que:

$$\Delta_{\boldsymbol{\gamma}} = \mathbf{Y}^\top \mathbf{K}_1 \left(\mathbf{A}_1^{1/2} \mathbf{R}(\phi) \mathbf{A}_1^{1/2} \right)^{-1} \text{diag}(\mathbf{z}^* - \boldsymbol{\pi}),$$

e

$$\Delta_{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{K}_2 \left(\mathbf{A}_2^{1/2} \mathbf{R}(\phi) \mathbf{A}_2^{1/2} \right)^{-1} \mathbf{Z}^* \text{diag}(\mathbf{z} - \boldsymbol{\mu}).$$

4.4.2.2 Esquema de perturbação da variável resposta do componente contínuo do modelo Quasi-gama aumentada com zero e espacialmente correlacionado

Tendo em vista que $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$, então, é permissível avaliar a influência local individual para os parâmetros $\boldsymbol{\gamma}, \boldsymbol{\beta}$. Portanto, aqui o interesse é verificar a influência que observações potencialmente influentes exercem sobre as estimativas $\boldsymbol{\beta}$. Desta forma, e tal qual é abordado na literatura, a curvatura normal C_d deve ser particionada, ou seja:

$$C_d(\boldsymbol{\beta}) = -2 \left(\mathbf{d}^\top \Delta^\top (\ddot{\mathcal{F}}^{-1} - \ddot{\mathcal{F}}^{\boldsymbol{\gamma}\boldsymbol{\gamma}}) \Delta \mathbf{d} \right),$$

em que

$$\ddot{\mathcal{F}}\gamma\gamma = \begin{pmatrix} \ddot{\mathcal{F}}\gamma\gamma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Não somente mas também, o gráfico de índice da matriz $\mathbf{B}(\beta) = \Delta^\top (\ddot{\mathcal{F}}^{-1} - \ddot{\mathcal{F}}\gamma\gamma)\Delta$ pode ser utilizado pra identificar as observações que exercem alguma influência nas estimativas de β . Nesta conjectura, seja $z_i > 0$ a variável resposta perturbada da seguinte forma $z_{i\omega_i} = z_i + \omega_i A_{2i}^{1/2}$, sendo $A_{2i}^{1/2}$ o i -ésimo elemento da matriz de função das variâncias. Então, com esta imposição, temos que $\omega_i = 0$, $i = 1, \dots, n$, indica ausência de perturbação na variável de interesse. Isto posto, a equação de estimativa para β do modelo perturbado fica dada por:

$$\mathbf{U}(\beta|\omega) = \mathbf{X}^\top \mathbf{K}_2 \left(\mathbf{A}_2^{1/2} \mathbf{R}(\phi) \mathbf{A}_2^{1/2} \right)^{-1} \mathbf{Z}^* (\mathbf{z}\omega - \mu).$$

Portanto, neste caso, temos que $\Delta\gamma = \mathbf{0}$. Por outro lado, a matriz $\Delta\beta$ é dada por:

$$\Delta\beta = -\mathbf{X}^\top \mathbf{K}_2 \left(\mathbf{A}_2^{1/2} \mathbf{R}(\phi) \mathbf{A}_2^{1/2} \right)^{-1} \mathbf{Z}^* \frac{\partial \mathbf{z}\omega}{\partial \omega^\top},$$

em que $\frac{\partial \mathbf{z}\omega}{\partial \omega^\top} = \text{diag}(\exp(\mathbf{x}^\top(\mathbf{s}_1)\beta), \dots, \exp(\mathbf{x}^\top(\mathbf{s}_n)\beta))$.

4.4.2.3 Esquema de perturbação simultânea nas variáveis preditoras do excesso de zeros e componente contínua

Para este caso, suponhamos que as matrizes das variáveis preditoras sejam distintas, ou seja $\mathbf{X} \neq \mathbf{Y}$, porém com as mesmas dimensões. Desta forma, consideramos uma perturbação do tipo aditiva tal que

$$y_{ij\omega_i} = y_{ij} + \sigma_y \omega_i,$$

$$x_{ij\omega_i} = x_{ij} + \sigma_x \omega_i,$$

em que σ_y e σ_x são constantes de escala dadas pelos desvios padrão da j -ésima coluna das matrizes $\mathbf{Y}$ e $\mathbf{X}$, respectivamente. Com isso, temos que

$$\zeta_{i\omega_i} = \gamma_1 y_{i1} + \dots + \gamma_j (y_{ij} + \sigma_y \omega_i) + \dots + \gamma_1 y_{ip} = \mathbf{y}_{\omega_i}^\top(\mathbf{s}_i) \gamma,$$

$$\eta_{i\omega_i} = \beta_1 x_{i1} + \dots + \beta_j (x_{ij} + \sigma_x \omega_i) + \dots + \beta_1 x_{ip} = \mathbf{x}_{\omega_i}^\top(\mathbf{s}_i) \beta,$$

em que $\mathbf{x}_{\omega_i}(\mathbf{s}_i) = (x_{i1}, \dots, x_{ij\omega_i}, \dots, x_{ip})^\top$, $\mathbf{y}_{\omega_i}(\mathbf{s}_i) = (y_{i1}, \dots, y_{ij\omega_i}, \dots, y_{ip})^\top$, bem como o vetor de não perturbação é $\omega_0 = (0, \dots, 0)^\top$. Então, com esta conjectura, as equações de

estimação, perturbada, são dadas por:

$$\begin{aligned} \mathbf{U}(\gamma|\boldsymbol{\omega}) &= \mathbf{Y}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{1\boldsymbol{\omega}} (\mathbf{V}_{1\boldsymbol{\omega}})^{-1} (\mathbf{z}^* - \boldsymbol{\pi}_{\boldsymbol{\omega}}), \\ \mathbf{U}(\beta|\boldsymbol{\omega}) &= \mathbf{X}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{2\boldsymbol{\omega}} (\mathbf{V}_{2\boldsymbol{\omega}})^{-1} \mathbf{Z}^* (\mathbf{z} - \boldsymbol{\mu}_{\boldsymbol{\omega}}), \end{aligned}$$

em que o índice $\boldsymbol{\omega}$ indica que as matrizes dependem da perturbação empregada. Por conseguinte, os elementos da matriz $\boldsymbol{\Delta}$ são calculados como segue:

$$\boldsymbol{\Delta}\gamma = \frac{\partial}{\partial \boldsymbol{\omega}^{\top}} \left\{ \mathbf{Y}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{1\boldsymbol{\omega}} (\mathbf{V}_{1\boldsymbol{\omega}})^{-1} \right\} (\mathbf{z}^* - \boldsymbol{\pi}_{\boldsymbol{\omega}}) - \mathbf{Y}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{1\boldsymbol{\omega}} (\mathbf{V}_{1\boldsymbol{\omega}})^{-1} \frac{\partial \boldsymbol{\pi}_{\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}^{\top}},$$

em que

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\omega}^{\top}} \left\{ \mathbf{Y}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{1\boldsymbol{\omega}} (\mathbf{V}_{1\boldsymbol{\omega}})^{-1} \right\} &= \frac{\partial \mathbf{Y}_{\boldsymbol{\omega}}^{\top}}{\partial \boldsymbol{\omega}^{\top}} \mathbf{K}_{1\boldsymbol{\omega}} (\mathbf{V}_{1\boldsymbol{\omega}})^{-1} + \mathbf{Y}_{\boldsymbol{\omega}}^{\top} \frac{\partial \mathbf{K}_{1\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}^{\top}} (\mathbf{V}_{1\boldsymbol{\omega}})^{-1} \\ &\quad + \mathbf{Y}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{1\boldsymbol{\omega}} \frac{\partial (\mathbf{V}_{1\boldsymbol{\omega}})^{-1}}{\partial \boldsymbol{\omega}^{\top}}, \end{aligned}$$

$\frac{\partial \mathbf{V}_{1\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}^{\top}} = \frac{\partial \mathbf{A}_{1\boldsymbol{\omega}}^{1/2}}{\partial \boldsymbol{\omega}^{\top}} \mathbf{R}(\phi) \mathbf{A}_{1\boldsymbol{\omega}}^{1/2} + \mathbf{A}_{1\boldsymbol{\omega}}^{1/2} \mathbf{R}(\phi) \frac{\partial \mathbf{A}_{1\boldsymbol{\omega}}^{1/2}}{\partial \boldsymbol{\omega}^{\top}}$. Ainda, os elementos de cada matriz de derivadas avaliadas em $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, são dados por:

$$\begin{aligned} \frac{\partial \mathbf{K}_{1\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}^{\top}} &= \frac{\partial}{\partial \boldsymbol{\omega}^{\top}} \left\{ \text{diag} \left(\frac{\exp(\mathbf{y}_{\omega_i}^{\top}(\mathbf{s}_i)\boldsymbol{\gamma})}{(1 + \exp(\mathbf{y}_{\omega_i}^{\top}(\mathbf{s}_i)\boldsymbol{\gamma}))^2} \right) \right\} \\ &= \sigma_y \gamma_j \text{diag} \left(\frac{\exp(\mathbf{y}_{\omega_i}^{\top}(\mathbf{s}_i)\boldsymbol{\gamma})(1 - \exp(\mathbf{y}_{\omega_i}^{\top}(\mathbf{s}_i)\boldsymbol{\gamma}))}{(1 + \exp(\mathbf{y}_{\omega_i}^{\top}(\mathbf{s}_i)\boldsymbol{\gamma}))^3} \right) \\ &= \sigma_y \gamma_j \text{diag} \left(\pi_i (1 - \pi_i)^2 \left(1 - \frac{\pi_i}{1 - \pi_i} \right) \right) \\ &= \sigma_y \gamma_j \text{diag} (\pi_i (1 - \pi_i) (1 - 2\pi_i)), \quad i = 1, \dots, n, \quad j = 1, \dots, p, \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathbf{A}_{1\boldsymbol{\omega}}^{1/2}}{\partial \boldsymbol{\omega}^{\top}} &= \frac{\partial}{\partial \boldsymbol{\omega}^{\top}} \left\{ [\pi_{i\omega_i} (1 - \pi_{i\omega_i})]^{1/2} \right\} \\ &= \text{diag} \left(\frac{1}{2} [\pi_{i\omega_i} (1 - \pi_{i\omega_i})]^{-1/2} \pi_{i\omega_i} (1 - \pi_{i\omega_i}) (1 - 2\pi_{i\omega_i}) \sigma_y \gamma_j \right) \\ &= \sigma_y \gamma_j \text{diag} \left(\frac{1}{2} [\pi_i (1 - \pi_i)]^{1/2} (1 - 2\pi_i) \right), \quad i = 1, \dots, n, \quad j = 1, \dots, p, \end{aligned}$$

$$\begin{aligned} \frac{\partial \boldsymbol{\pi}_{\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}^{\top}} &= \frac{\partial}{\partial \boldsymbol{\omega}^{\top}} \left\{ \text{diag} \left(\frac{\exp(\mathbf{y}_{\omega_i}^{\top}(\mathbf{s}_i)\boldsymbol{\gamma})}{(1 + \exp(\mathbf{y}_{\omega_i}^{\top}(\mathbf{s}_i)\boldsymbol{\gamma}))} \right) \right\} \\ &= \sigma_y \gamma_j \text{diag} \left(\pi_i (1 - \pi_i)^2 \left(1 - \frac{\pi_i}{1 - \pi_i} \right) \right) \\ &= \sigma_y \gamma_j \text{diag} (\pi_i (1 - \pi_i)), \quad i = 1, \dots, n, \quad j = 1, \dots, p, \end{aligned}$$

enquanto que $\frac{\partial \mathbf{Y}_{\boldsymbol{\omega}}^{\top}}{\partial \boldsymbol{\omega}^{\top}}$ é uma matriz $p \times n$ de zeros exceto a p -ésima linha que é preenchida com a constante σ_y . Analogamente,

$$\boldsymbol{\Delta}\beta = \frac{\partial}{\partial \boldsymbol{\omega}^{\top}} \left\{ \mathbf{X}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{2\boldsymbol{\omega}} (\mathbf{V}_{2\boldsymbol{\omega}})^{-1} \right\} \mathbf{Z}^* (\mathbf{z} - \boldsymbol{\mu}_{\boldsymbol{\omega}}) - \mathbf{X}_{\boldsymbol{\omega}}^{\top} \mathbf{K}_{2\boldsymbol{\omega}} (\mathbf{V}_{2\boldsymbol{\omega}})^{-1} \mathbf{Z}^* \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}^{\top}},$$

em que

$$\begin{aligned}
 \frac{\partial \mathbf{K}_{2\omega}}{\partial \omega^\top} &= \sigma_x \beta_j \text{diag} \left(\exp(\mathbf{x}_{\omega_i}^\top (s_i) \beta) \right) \\
 &= \sigma_x \beta_j \text{diag} (\mu_i), \quad i = 1, \dots, n, \quad j = 1, \dots, p, \\
 &= \frac{\partial \boldsymbol{\mu}_\omega}{\partial \omega^\top} = \frac{\partial \mathbf{A}_{2\omega}^{1/2}}{\partial \omega^\top},
 \end{aligned}$$

enquanto que $\frac{\partial \mathbf{X}_\omega^\top}{\partial \omega^\top}$ é uma matriz $p \times n$ de zeros exceto a p -ésima linha que é preenchida com a constante σ_x .

4.5 ESTUDO DE SIMULAÇÃO

Nesta Seção aplicamos um estudo de simulação para avaliar o desempenho do algoritmo ES sob GEE para o modelo Quasi-gama aumentada com zero e espacialmente correlacionado. Com isso, concentramos na investigação das estimativas $\hat{\gamma}$, $\hat{\beta}$ e na eficiência das medidas de influência local na identificação de observações potencialmente influentes. Para tanto, o modelo é considerado com a seguinte estrutura:

$$\begin{aligned}
 h_1(\pi_i) &= \gamma_0 + \gamma_1 y_1(s_i) + \gamma_2 y_2(s_i) \\
 h_2(\mu_i) &= \beta_0 + \beta_1 x_1(s_i) + \beta_2 x_2(s_i),
 \end{aligned}$$

em que $i = 1, \dots, n$, com matrizes de covariáveis preditoras dadas por $\mathbf{X} \sim U(0, 1)$ e $\mathbf{Y} \sim U(0; 0, 5)$, sendo todas as covariáveis fixadas ao longo das simulações. Ainda, temos que h_1 e h_2 são as funções de ligação *logito* e *logarítmica*, respectivamente. Com isso, a variável resposta z_i é gerada pelo algoritmo 3.

Algorithm 3 Geração de dados com distribuição quasi-gama aumentada com zero e espacialmente correlacionada.

Passo 1 - Seja π_i a probabilidade de ocorrência do valor zero observado. Então, se a amostra aleatória $\text{Binomial}(1, 1, \pi_i) = 1$, defina $z_i = 0$;

Passo 2 - Caso contrário, determine uma matriz não singular de correlação espacial Σ ;

Passo 3 - Gere um vetor aleatório $\mathbf{Z}_1$, tal que $\mathbf{Z}_1 \sim MVN(\mathbf{0}, \Sigma)$;

Passo 4 - Determine o vetor aleatório $\mathbf{Z}$, tal que $z_i \mapsto F^{-1}(\Phi(z_{1_i}))$, $i = 1, \dots, n$;

Passo 4 - Para cada z_i calcule a distribuição gamma inversa (quantil).

O algoritmo 3 foi construído a partir do método de geração de variáveis correlacionadas *NORTA* proposto por Cario e Nelson (1997). Ainda, consideramos três cenários para o tamanho

da amostra, a saber: com $n = 100$, $n = 200$ e $n = 300$, com pontos amostrais gerados em grades regulares de dimensões 10×10 , 20×10 e 15×20 , respectivamente, considerando um espaçamento de 1 (uma) unidade entre as observações. Os parâmetros do modelo de regressão são fixados em $\beta = (0, 1; 0, 1; 0, 1)^\top$ e $\gamma = (0, 5; 0, 5; 0, 5)^\top$. Para a estrutura de dependência espacial foram utilizadas as famílias de covariância Wendland e Matérn, com suavização $\kappa = 2$ e $\nu = 1, 5$, respectivamente. O vetor paramétrico que define as famílias de covariâncias foi fixado em $\phi = (0, 1; 0, 6; 1)^\top$. O número de réplicas de Monte Carlo foi de $R = 1000$, em que foram gerados um vetor de resposta $\mathbf{z} = (z_1, \dots, z_n)^\top$, para cada uma das réplicas. As estimativas dos parâmetros de regressão foram obtidas pelo algoritmo ES (2). Para a análise dos resultados, foram calculados o $\hat{E}(\hat{\theta}) = R^{-1} \sum_{i=1}^R \hat{\theta}_i$, o viés $\hat{B}(\hat{\theta}) = \hat{E}(\hat{\theta}) - \theta$, a raiz quadrada do erro quadrático médio $\sqrt{EQM} = \sqrt{R^{-1} \sum_{i=1}^R (\hat{\theta}_i - \theta)^2}$, em que θ é o vetor de parâmetros verdadeiro e $\hat{\theta}$ é o estimador. Todo o procedimento de cálculo foi realizado utilizando o *software* R Core Team (2020).

Nas Tabelas (21), (22), (23), (24), (25) e (26) apresentamos os resultados das simulações para a distribuição quasi-gama aumentada com zero e espacialmente correlacionada. De modo geral, as estimativas obtidas para o parâmetro β apresentam um bom desempenho com relação ao viés e $\sqrt{EQM}$, mesmo para amostras pequenas. Por outro lado, as estimativas para o parâmetro γ são bem enviesadas, além de apresentar, dentre as 1000, valores de estimativas muito distantes do verdadeiro parâmetro, o que pode ser visto pelo comportamento dos valores de $\sqrt{EQM}$ nas Tabelas (21), (22), (25) e (26). Em síntese, o estimador para modelar o parâmetro da parte contínua mostrou-se mais eficiente que o estimador para o excesso de zeros, este último sendo bastante viesado.

Paralelamente, ilustramos o uso da técnica de influência local na detecção de possíveis observações influentes, considerando os esquemas de perturbação do tipo ponderação de casos, na resposta da parte contínua e em uma das covariáveis tanto da parte do excesso de zeros, quanto da parte contínua. Após gerar os dados, perturbamos o valor máximo de cada uma das 1000 réplicas para gerar um ponto discrepante, tal que $z_{max} \leftarrow z_{max} + a\sqrt{\mathbf{z}^\top \mathbf{z}}$, em que $a \in \{0; 0, 1; 0, 25; 0, 5; 0, 75; 1; 2; 4\}$ é uma constante de escala. Com isso, ajustamos o modelo quasi-gama aumentado com zero espacialmente correlacionado e calculamos o vetor $\mathbf{C}_{d_{max}}$ para $\theta = (\gamma^\top, \beta^\top)^\top$, com $(C_i = 2|b_{ii}|)$, em que b_{ii} é o i -ésimo elemento da diagonal da matriz $\mathbf{B} = -\Delta^\top \ddot{\mathcal{F}}^{-1} \Delta$, $i = 1, \dots, n$. Aqui, consideramos a i -ésima observação como potencialmente influente se $C_i > 2 * sd(\mathbf{C}_{d_{max}})$. A partir disto, um contador é adicionado para contabilizar o percentual de vezes em que a observação discrepante é identificada, sob

cada esquema de perturbação. Nas Tabelas 27 e 28, apresentamos o percentual de detecção do ponto perturbado. Os resultados mostram que a medida de influência local é eficiente na identificação de observações discrepantes, com destaque para o esquema de perturbação nas covariáveis. Portanto, os resultados evidenciam a utilidade efetiva, da técnica de influência local, para identificação de observações potencialmente influentes.

Tabela 21 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Wendland ($\kappa = 2$) e suas respectivos $\sqrt{EQM}$ e viés.

Amostra de tamanho $n = 100$.							
β	$\hat{\beta}$	Viés	$\sqrt{EQM}$	γ	$\hat{\gamma}$	viés	$\sqrt{EQM}$
0,100	0,107	0,007	0,210	0,500	0,666	0,166	0,922
0,100	0,109	0,009	0,298	0,500	0,144	-0,356	0,761
0,100	0,107	0,007	0,214	0,500	0,140	-0,360	0,720

Fonte: Elaborada pelo autor (2022).

Tabela 22 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Matérn ($\nu = 1,5$) e suas respectivos $\sqrt{EQM}$ e viés.

Amostra de tamanho $n = 100$.							
β	$\hat{\beta}$	Viés	$\sqrt{EQM}$	γ	$\hat{\gamma}$	viés	$\sqrt{EQM}$
0,100	0,106	0,006	0,161	0,500	0,510	0,010	0,939
0,100	0,110	0,010	0,224	0,500	0,072	-0,428	0,867
0,100	0,109	0,009	0,212	0,500	0,148	-0,352	0,835

Fonte: Elaborada pelo autor (2022).

Tabela 23 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação Wendland ($\kappa = 2$) e suas respectivos $\sqrt{EQM}$ e viés.

Amostra de tamanho $n = 200$.							
β	$\hat{\beta}$	Viés	$\sqrt{EQM}$	γ	$\hat{\gamma}$	viés	$\sqrt{EQM}$
0,100	0,106	0,006	0,110	0,500	0,547	0,047	0,360
0,100	0,134	0,034	0,148	0,500	0,111	-0,389	0,552
0,100	0,131	0,031	0,138	0,500	0,138	-0,362	0,540

Fonte: Elaborada pelo autor (2022).

Tabela 24 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Matérn ($\nu = 1,5$) e suas respectivos $\sqrt{EQM}$ e viés.

Amostra de tamanho $n = 200$.

β	$\hat{\beta}$	Viés	$\sqrt{EQM}$	γ	$\hat{\gamma}$	viés	$\sqrt{EQM}$
0,100	0,108	0,008	0,105	0,500	0,544	0,044	0,346
0,100	0,133	0,033	0,138	0,500	0,120	-0,380	0,547
0,100	0,127	0,027	0,138	0,500	0,134	-0,366	0,546

Fonte: Elaborada pelo autor (2022).

Tabela 25 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Wendland ($\kappa = 2$) e suas respectivos $\sqrt{EQM}$ e viés.

Amostra de tamanho $n = 300$.

β	$\hat{\beta}$	Viés	$\sqrt{EQM}$	γ	$\hat{\gamma}$	viés	$\sqrt{EQM}$
0,100	0,105	0,005	0,149	0,500	0,625	0,125	1,432
0,100	0,109	0,009	0,351	0,500	0,191	-0,309	0,501
0,100	0,118	0,018	0,566	0,500	0,177	-0,323	0,531

Fonte: Elaborada pelo autor (2022).

Tabela 26 – Estimativas para os parâmetros de regressão sob GEE das 1000 réplicas de simulação do modelo quasi-gama aumentado com zero e espacialmente correlacionado com matriz de correlação família Matérn ($\nu = 1,5$) e suas respectivos $\sqrt{EQM}$ e viés.

Amostra de tamanho $n = 300$.

β	$\hat{\beta}$	Viés	$\sqrt{EQM}$	γ	$\hat{\gamma}$	viés	$\sqrt{EQM}$
0,100	0,107	0,007	0,094	0,500	0,752	0,252	1,279
0,100	0,107	0,007	0,131	0,500	0,192	-0,308	0,491
0,100	0,117	0,017	0,152	0,500	0,195	-0,305	0,526

Fonte: Elaborada pelo autor (2022).

Tabela 27 – Percentual de vezes em que a observação perturbada (caracterizada como influente) foi detectada sobre as 1000 réplicas, sob os esquema de perturbação ponderação de casos, variável resposta e covariáveis. A matriz de covariância que gerou os dados foi a Wendland ($\kappa = 2$). Considerou-se a medida de influência local total, correspondente a direção da i -ésima observação ($C_i = 2|b_{ii}|$).

a	$n = 100$			$n = 200$			$n = 300$		
	P.C	P.Res	P.Cov	P.C	P.Res	P.Cov	P.C	P.Res	P.Cov
0,00	87,7	81,7	95,0	90,2	75,0	99,2	89,4	64,9	99,4
0,10	88,7	83,6	98,9	90,6	76,2	99,9	90,5	65,6	99,7
0,25	89,7	86,1	99,8	91,1	78,5	100	92,2	64,2	99,9
0,50	92,2	87,4	99,9	92,2	79,5	99,9	93,2	62,2	99,9
0,75	93,2	87,7	100	91,3	80,5	100	93,1	59,5	99,9
1,00	93,2	87,1	100	90,3	79,6	100	91,2	56,2	99,9
2,00	90,5	83,6	100	85,5	73,4	100	83,7	43,6	98,4
4,00	87,9	72,9	99,1	83,9	60,0	99,6	82,4	28,1	85,8

Fonte: Elaborada pelo autor (2022).

Esquemas de perturbação: P.C - Ponderação de casos. P.Res - Variável resposta. P.Cov - Covariáveis contínua e discreta. a = constante de escala.

Tabela 28 – Percentual de vezes em que a observação perturbada (caracterizada como influente) foi detectada sobre as 1000 réplicas, sob os esquema de perturbação ponderação de casos, variável resposta e covariáveis. A matriz de covariância que gerou os dados foi a Matérn ($\nu = 1, 5$). Considerou-se a medida de influência local total, correspondente a direção da i -ésima observação ($C_i = 2|b_{ii}|$).

a	$n = 100$			$n = 200$			$n = 300$		
	P.C	P.Res	P.Cov	P.C	P.Res	P.Cov	P.C	P.Res	P.Cov
0,00	80,8	79,0	97,2	91,0	84,0	99,4	90,4	84,9	99,8
0,10	84,3	81,6	99,2	92,6	87,0	99,6	92,3	87,0	99,8
0,25	89,4	85,0	99,8	93,3	85,9	99,8	93,8	90,1	99,8
0,50	94,6	91,2	100	72,8	56,4	79,3	96,0	93,4	100
0,75	96,4	95,1	100	77,3	54,0	79,2	98,2	95,6	100
1,00	96,8	97,5	100	79,7	52,9	79,1	98,7	96,7	100
2,00	93,8	99,5	100	84,7	46,5	78,6	95,6	98,2	100
4,00	91,6	98,2	99,8	83,4	40,1	78,4	92,3	96,3	100

Fonte: Elaborada pelo autor (2022).

Esquemas de perturbação: P.C - Ponderação de casos. P.Res - Variável resposta. P.Cov - Covariáveis contínua e discreta. a = constante de escala.

4.6 APLICAÇÃO: CHUVA ACUMULADA (MM) DO MÊS DE AGOSTO DE 2021 NO ESTADO DE PERNAMBUCO (BR)

O objetivo desta Seção é ilustrar a metodologia desenvolvida neste Capítulo no que concerne aos estimadores dos parâmetros, e análise de diagnóstico de influência local, para o

modelo Quasi-gama aumentada com zero e espacialmente correlacionado. Para tanto, consideramos um conjunto de dados referente a Chuva acumulada (mm) do mês de agosto de 2021 no estado de Pernambuco (BR). As observações correspondem a 190 pontos georreferenciados com proporção de zeros igual a 13,68%, coletados pela Agência Pernambucana de Águas e Clima (APAC). Com isso, o interesse é modelar a variável z que denota a precipitação pluviométrica acumulada para o período considerado. As variáveis explicativas são altitude (m) e temperaturas máxima $^{\circ}C$. Então, inicialmente é realizada uma análise descritiva das variáveis envolvidas, ver Tabela 29. Desta forma, temos que a variável resposta assume valores entre 0 e 478,1mm com distribuição assimétrica (ver Figura 21(a)). Ainda, notamos uma alta dispersão com uma alta concentração de observações próximas do zero, significando que a quantidade de chuva nestas localizações é relativamente baixa. No gráfico de histograma o destaque vertical no ponto zero informa a quantidade de zeros da amostra. Por outro lado, no gráfico de *box-plot* notamos a presença de valores atípicos, os quais são (385, 1; 367; 478, 1; 378, 5; 459, 7; 397, 1; 428; 373, 7; 412, 4) correspondentes as observações 1, 3, 4, 5, 10, 21, 48, 54 e 55.

Na Figura 21 temos os gráficos de *post-plot* segundo o quartil dos dados e semivariograma empírico. Da análise do gráfico Figura 22(a) verificamos que os maiores valores aglomeram-se em estações do litoral do estado. De outra forma, a Figura 22 ilustra a dispersão, correlação linear e distribuição das variáveis envolvidas no estudo. De sua análise, notamos uma correlação negativa de ambas as covariáveis com a resposta. Com isso, e a partir dos gráficos de dispersão, percebemos que quando as covariáveis crescem a resposta tende a diminuir.

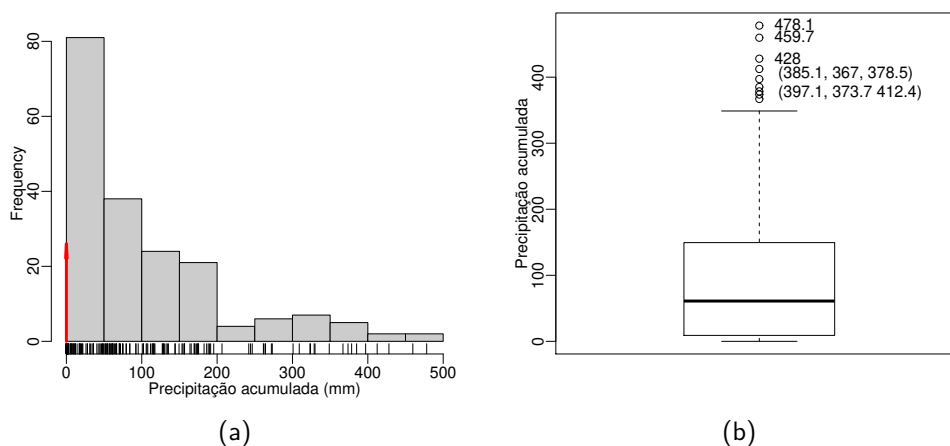
Tabela 29 – Medidas resumo para variável precipitação pluviométrica acumulada (Prec).

Variável	n	Min.	Q_1	Mediana	Média	Q_3	Max.	Var	SD	CV
Prec	190	0,0	9,0	61,0	97,78	148,25	478,1	12040,25	109,73	112,2
Altitude	190	8,0	160,0	437,5	401,0	555,5	1010,0	58586,46	242,05	60,37
Max.Temp	190	23,0	25,5	26,5	26,71	28,7	30,6	4,01	2,0	7,5

Fonte: Elaborada pelo autor (2022).

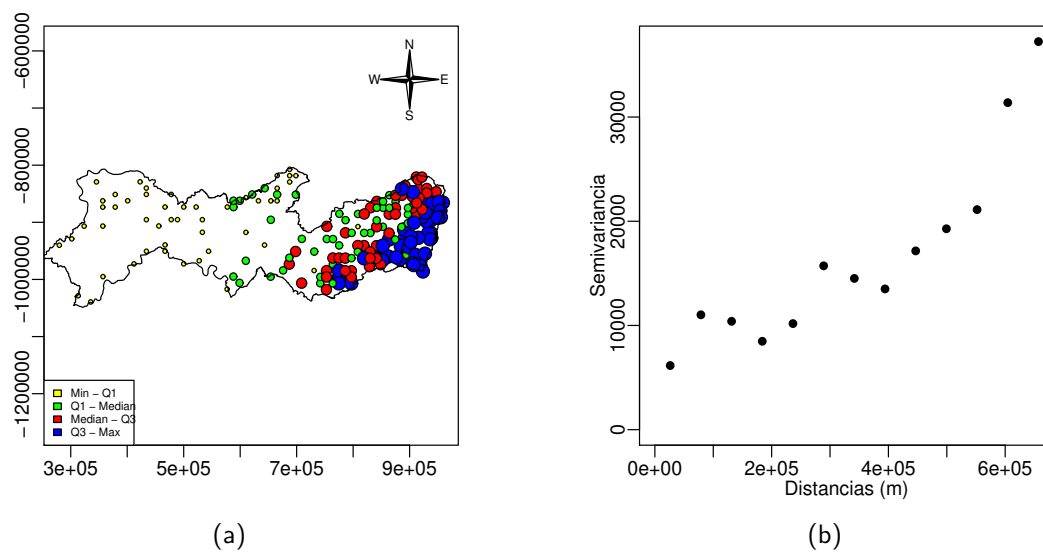
Min.: mínimo, 1st Q_1 : primeiro quartil, Q_3 : terceiro quartil, Max.: máximo, SD: Desvio padrão, Var: Variância, Cv: Coeficiente de variação

Figura 20 – Histograma de frequências e *BoxPlot* da variável precipitação pluviométrica acumulada (*mm*) o mês de agosto de 2021 no estado de Pernambuco, Brasil, Figuras 21(a) e 21(b), respectivamente.



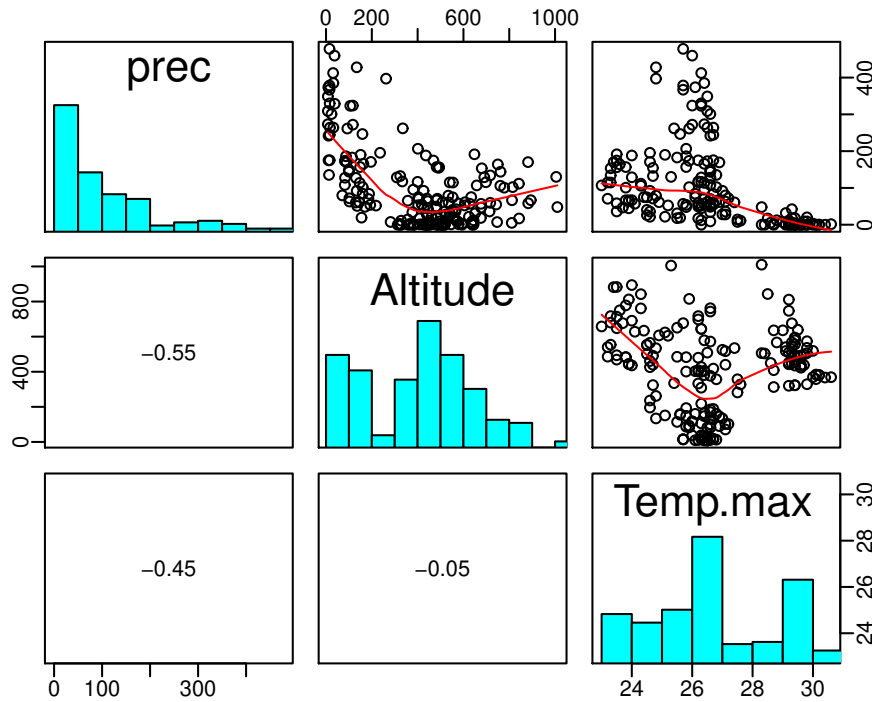
Fonte: Elaborada pelo autor (2022).

Figura 21 – Gráfico *post-plot* segundo o quartil dos dados (Figura 22(a)) e semivariograma empírico da variável precipitação pluviométrica acumulada (Figura 22(b)).



Fonte: Elaborada pelo autor (2022).

Figura 22 – Gráfico de dispersão e correlação linear de pearson das variáveis envolvidas no estudo.



Fonte: Elaborada pelo autor (2022).

Outrossim, para este conjunto de dados assumimos um modelo de regressão quasi-gama aumentado com zero e espacialmente correlacionado, tal que:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \gamma_0 + \sum_{j=1}^p \gamma_j y_j(s_i)$$

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j x_j(s_i),$$

em que os parâmetros π_i e μ_i representam a probabilidade da variável resposta assumir o valor zero e a média, respectivamente, enquanto que $y_j(s_i)$ e $x_j(s_i)$ são os elementos correspondentes a j -ésima coluna das matrizes de covariáveis.

Com isso, dado o vetor paramétrico $\phi = (\phi_1, \phi_2, \phi_3)^\top$, a matriz de correlação espacial é estimada e, então, introduzida como constante no algoritmo ES 3, assim, estimamos os parâmetros do excesso de zeros e da média. Os parâmetros iniciais para o processo iterativo são obtidos assumindo independência e estimados pela função *gamlss* do pacote desenvolvido por (STASINOPOULOS; RIGBY et al., 2007) no *software* R Core Team (2020).

Destarte, a partir da análise do gráfico de semivariograma (Figura 22(b)), e considerando a variância dos dados, fixamos o vetor da estrutura da dependência espacial em $(\phi_1, \phi_2, \phi_3)^\top = (6500, 12500, 100)^\top$. Os parâmetros iniciais para o modelo do componente contínuo e discreto, foram obtidos supondo independência, sendo iguais a $\gamma = (-120, 321; 0, 244; 35, 077)^\top$ e $\beta = (47, 658; -0, 576; -12, 288)^\top$. Para explicar a correlação entre as observações utilizamos uma função de pertencente a família Wendland com suavização $\kappa = 2$. Assim, foram realizados vários ajustes, então, a partir da análise dos resíduos, bem como a qualidade do ajuste a partir do gráfico normal de probabilidade com envelopes simulados, retiramos as observações de ordem 73, 123, 169 da análise por serem consideradas discrepantes, assim escolhemos o modelo:

$$\begin{aligned} \log\left(\frac{\pi_i}{1 - \pi_i}\right) &= \gamma_0 + \gamma_1 \log(\text{Altitude}) + \gamma_2 \log(\text{Temp.max}) \\ \log(\mu_i) &= \beta_0 + \beta_1 \log(\text{Altitude}) + \beta_2 \log(\text{Temp.max}) + \log(M.\text{Clima}), \end{aligned}$$

em que $\log(M.\text{Clima})$ é a variável (*offset*) utilizada para incorporar a heterogeneidade da quantidade de chuva na região em estudo, com M.Clima denotando a média climatológica dos últimos 30 anos.

A Tabela 30 apresenta as estimativas dos parâmetros, tal como os respectivos erros padrão assintóticos robustos. Observamos que, para o parâmetro do componente contínuo, variações no intercepto causa um aumento na média da variável resposta dos valores positivos. Ocorrendo estritamente o oposto para as variáveis logaritmos da Altitude e temperatura máxima. Contrariamente, as estimativas para o componente discreto indicam que, variações no intercepto e logaritmos da Altitude, provocam uma diminuição na probabilidade de ocorrência de zero. Enquanto que, uma variação na variável logaritmo da temperatura máxima causa um aumento na probabilidade de uma observação ser zero. Com relação aos erros padrão, os resultados foram similares para ambos os submodelos, ocorrendo valores maiores para o intercepto.

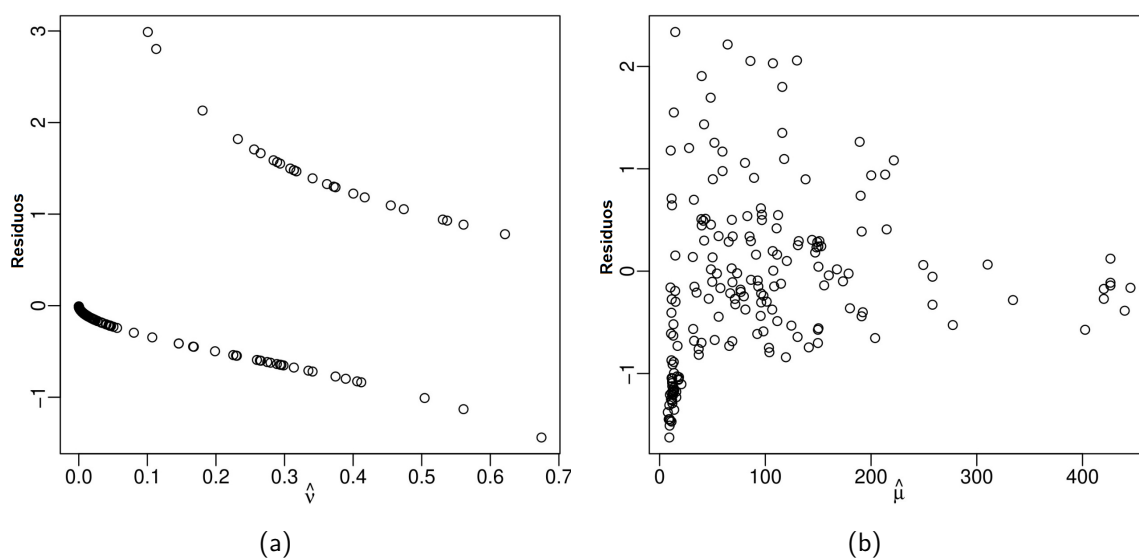
Tabela 30 – Estimativas dos parâmetros de regressão para o modelo quasi-gama aumentado com zero e espacialmente correlacionado pelo algoritmo *Expectation-Solution* (ES).

Parâmetros	Estimativas	Erro-padrão robusto
$\hat{\beta}_0$	47,814	2,727
$\hat{\beta}_1$	-0,579	0,037
$\hat{\beta}_2$	-12,383	0,699
$\hat{\gamma}_0$	-119,958	2,722
$\hat{\gamma}_1$	-0,359	0,034
$\hat{\gamma}_2$	35,899	0,745

Fonte: Elaborada pelo autor (2022).

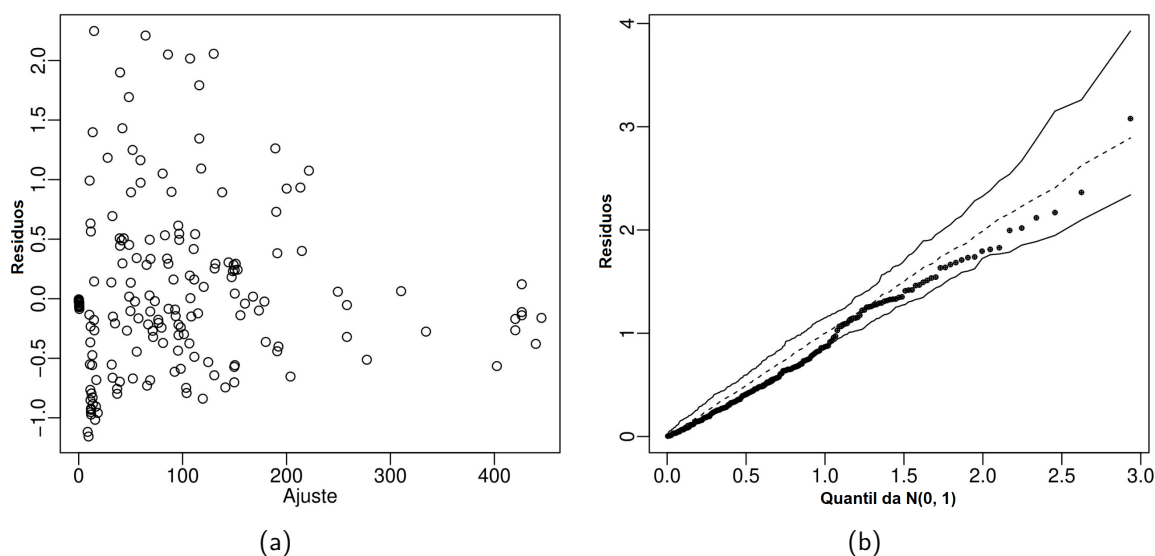
Finalizando esta primeira parte de diagnóstico, nas figuras 24(a), 24(b), 25(a) e 25(b), apresentamos os gráficos dos resíduos contra a média do componente discreto, contínuo, modelo completo contra os valores ajustados e, gráfico normal de probabilidade com envelopes simulados para o modelo ajustado, respectivamente. Analisando os resíduos notamos que os pontos estão distribuídos dispersos aleatoriamente em torno de zero, concentrados entre -2 e 2 , indicando, assim, que o modelo está bem ajustado. O gráfico de envelope simulado confirma a conclusão da qualidade do ajuste, visto que todas as observações estão dentro ou sobre os limites determinados pelo envelope.

Figura 23 – Resíduos para o modelo dos componentes discreto e contínuo, Figuras 24(a) e 24(b), respectivamente, contra seus respectivos valores ajustados para os dados de precipitação pluviométrica acumulada.



Fonte: Elaborada pelo autor (2022).

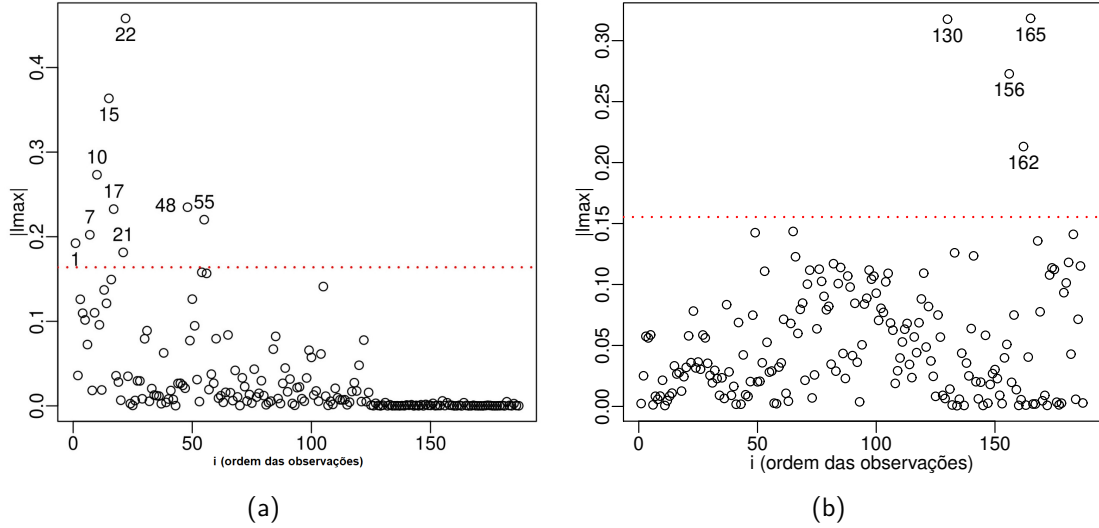
Figura 24 – Resíduos para o modelo completo, Figura 25(a), contra os valores ajustados. Gráfico normal de probabilidade com envelopes simulados para o modelo ajustado, Figura 25(b), para os dados de precipitação pluviométrica acumulada.



Fonte: Elaborada pelo autor (2022).

Finalmente, com o objetivo de verificar a existência de pontos potencialmente influentes para as estimativas dos parâmetros, aplicamos a técnica de influência local. Para identificar estas observações, consideramos os gráficos de índices de $|l_{max}|$ (autovetor associado ao maior autovalor relacionado a maior curvatura normal) sob um esquema de perturbação ponderação de casos, Figura 26(a), e perturbação nas respostas do componente contínuo, Figura 26(b). As observações foram consideradas potencialmente influente, caso fossem maiores que a média mais duas vezes o desvio padrão de $|l_{max}|$ (ver linha horizontal pontilhada). No esquema de ponderação de casos, identificamos as observações de ordens 1, 7, 10, 15, 17, 21, 22, 48 e 55 como potencialmente influentes, enquanto para perturbação nas respostas, identificamos as observações 130, 156, 162 e 165.

Figura 25 – Gráfico de índices de $|lmax|$ (autovetor associado ao maior autovalor relacionado a maior curvatura normal) para ponderação de casos, Figura 26(a), e perturbação nas respostas, Figura 26(b), para os dados de precipitação pluviométrica acumulada.



Fonte: Elaborada pelo autor (2022).

Com vistas na Figura 22(a) (post-plot segundo os quartis dos dados), as observações classificadas nos gráficos 26(a) e 26(b) como potencialmente influente não apresentam variações (com relação a sua magnitude) discrepantes, em comparação aos seus vizinhos mais próximos. Contudo, o gráfico de box-plot (Figura 21(b)), identifica as observações 1, 10, 21, 48 e 55 como outliers, tal qual o gráfico de influência 26(a), com relação a ponderação de casos. Diante disso, optamos por retirá-las conjuntamente e individualmente do conjunto de dados. Feito isto, a análise é novamente realizada, permitindo assim verificar o impacto destas observações nas estimativas dos parâmetros dos submodelos, consequentemente na qualidade do ajuste. Com isso, para avaliar a magnitude da influência nas estimativas dos parâmetros, utilizamos o desvio percentual relativo (*Desviopercentualrelativo(DPR)*), calculado por

$$DPR = \frac{\hat{\theta}_1 - \hat{\theta}}{\hat{\theta}} \times 100,$$

em que $\hat{\theta}_1$ é o estimador do vetor paramétrico θ quando retirada uma ou mais observações. Por outro lado, a influência no ajuste do modelo é mensurada pelo desvio quadrático médio relativo (Desvio quadrático médio relativo (DQMR)), calculado por:

$$DQMR = \frac{\sqrt{DQM_1} - \sqrt{DQM}}{\sqrt{DQM}} \times 100,$$

em que $DQM = \frac{\sum_{i=1}^n (z_i - \hat{z}_i)^2}{n}$ e DQM_1 é o DQM quando se retirada uma ou mais observações. Os resultados constam na Tabela 31.

Tabela 31 – Estimativas e desvio percentual relativo para os parâmetros dos submodelos quando se é retirada observações potencialmente influentes. Erros padrão encontram-se entre parênteses.

Obs. retirada	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	DQMR
Completo	-119,958 (2,722)	-0,359 (0,034)	35,899 (0,745)	47,814 (2,727)	-0,579 (0,037)	-12,383 (0,699)	— — —
1, 10, 21, 48, 55	-119,560 (2,748)	-0,360 (0,033)	35,798 (0,754)	46,482 (2,630)	-0,553 (0,035)	-12,031 (0,676)	-18,687
	-0,298%	0,255%	-0,281%	-2,786%	-4,515%	-2,849%	
1	-119,931 (2,721)	-0,360 (0,034)	35,892 (0,745)	47,796 (2,729)	-0,572 (0,036)	-12,390 (0,700)	-2,249
	-0,022%	0,151%	-0,019%	-0,037%	-1,103%	0,057%	
10	-119,959 (2,720)	-0,359 (0,034)	35,898 (0,744)	47,671 (2,724)	-0,571 (0,036)	-12,354 (0,699)	-3,752
	0,001%	-0,225%	-0,002%	-0,299%	-1,332%	-0,237%	
21	-119,856 (2,737)	-0,360 (0,033)	35,872 (0,750)	47,275 (2,678)	-0,580 (0,037)	-12,221 (0,687)	-4,310
	-0,085%	0,266%	-0,076%	-1,127%	0,220%	-1,313%	
48	-119,783 (2,737)	-0,359 (0,033)	35,849 (0,750)	47,342 (2,696)	-0,575 (0,036)	-12,246 (0,692)	-3,620
	-0,146%	-0,011%	-0,140%	-0,987%	-0,578 %	-1,108%	
55	-119,902 (2,722)	-0,360 (0,034)	35,884 (0,745)	47,741 (2,726)	-0,572 (0,036)	-12,374 (0,699)	-2,635
	-0,046%	0,117%	-0,043%	-0,153%	-1,259%	-0,074%	

Fonte: Elaborada pelo autor (2022).

Sendo assim, analisando os resultados constantes na Tabela 31, notamos que as observações apenas exerceram influência significativa, nas estimativas dos parâmetros do submodelo do componente contínuo (β), quando foram retiradas conjuntamente do conjunto de dados, desde que consideramos um (DPR) superior a 2%. Por outro lado, para o impacto no ajuste do modelo, faz sentido considerar influência significativa quando o valor de $DQMR$ for superior a 5% ou 10%, pois neste caso, a medida é aplicada sobre o ajuste do modelo completo. Portanto, analisando os resultado percebemos um impacto, novamente, apenas quando as observações são restiradas em conjunto. Portanto, concluímos que a modelagem da precipitação pluviométrica acumulada deve ser realizada com cautela sob a presença de observações potencialmente influentes.

4.7 CONSIDERAÇÕES FINAIS DO CAPÍTULO

A abordagem deste Capítulo propões a modelagem conjunta do excesso de zeros e da dependência espacial em modelos geoestatísticos. A proporção de zeros em excesso dificulta a obtenção de informações, sobretudo quando as observações diferentes de zero não provêm de uma distribuição Gaussiana. Com isso, e tendo como motivação dados meteorológicos no estado de Pernambuco, Brasil, este Capítulo apresentou-se como uma importante ferramenta no sentido de contribuir com uma proposta relevante para o contexto da análise de dados com estas características, e que são tipicamente assimétricos. Desenvolvemos a modelagem de dados contínuos positivos com a inclusão de zeros, a partir de um modelo particionado por uma regressão logística e uma função de quasi-verossimilhança da gama. Esta metodologia foi desenvolvida com a suposição da variável resposta assumir uma distribuição de mistura Bernoulli degenerada em zero e gama. Desta forma, o modelo permitiu explicar a variabilidade dos dados, inclusive o excesso de zeros. Além disso, o modelo permitiu incorporar a correlação espacial presente nos dados por meio das equações de estimação generalizadas, que retornam estimadores consistentes, eficientes e assintoticamente normais. Sendo assim, com esta conjectura, o modelo proposto permite que a matriz de covariância espacial seja estimada a partir dos dados, ou pressuposta de alguma classe de covariância espacial existente na literatura.

Outrossim, para o modelo apresentado, propomos a estimação dos parâmetros utilizando preditores lineares com funções de ligação adequadas, e à estimamos por processo iterativo com o algoritmo Expectation-Solution (ES). Ademais, com o objetivo de avaliar as pressuposições do modelo, desenvolvemos medidas de influência local a partir da curvatura normal, que por um estudo de simulação verificamos o seu potencial em identificar observações potencialmente influentes. E, por fim, realizamos um aplicação como exemplo numérico utilizando dados atuais de grande importância.

5 CONCLUSÕES GERAIS

Como é bem discutido na literatura, para a análise de diversos fenômenos utilizando a estatística clássica não é considerado qualquer conhecimento da posição geográfica das amostras, muito menos o relacionamento entre elas. Esta limitação é resolvida introduzindo os conceitos da geoestatística, na qual da forma como são definidos leva em consideração além da posição geográfica da amostra também o relacionamento entre elas. Diante disso, o objetivo deste estudo foi de reunir as principais ferramentas da geoestatística e propor novas metodologias de análise de dados, com aplicações no cotidiano, sob o enfoque de distribuições Gaussianas e não-gaussianas para a resposta.

Desta maneira, foram obtidos importantes resultados. Na Seção 2 o enfoque foi apresentar as funções de covariâncias pertencentes a família Wendland, e verificar o desempenho de sua aplicabilidade em modelos lineares espaciais Gaussianos. Com isso, comparamos os resultados com a conhecida família Matérn. Seguindo este viés, obtemos resultados ainda não disponíveis na literatura das expressões para o Índice de Dependência Espacial (IDE) e Área de Continuidade Espacial (ACE) para as classes Wendland e Matérn, além de expressões para alcance prático da família Wendland. Ainda, foram estendidas algumas medidas de influência local para o modelo linear espacial Gaussiano quando a covariância é um membro da família Wendland. Propomos níveis de referência, ou pontos de corte, para classificar observações como potencialmente influentes a partir do algoritmo Jackknife-after-Bootstrap. Como resultado, a partir de um exemplo numérico com dados de rendimento de soja, os pontos classificados como influentes afetaram as estimativas dos parâmetros que definem a estrutura de dependência espacial e tiveram influência na construção dos mapas de previsão.

Na Seção 3, propomos um novo critério para seleção da matriz de correlação de trabalho em modelos espaciais de quasi-verossimilhança. Realizamos o processo de estimação dos parâmetros de interesse por Equações de Estimação Generalizadas, no qual foi possível dilatar a suposição da distribuição dos dados para além da normal, considerando, assim, distribuições de probabilidade pertencentes à família exponencial. Importantes medidas de influência local foram desenvolvidas a partir do distanciamento das quasi-verossimilhanças, para as distribuições marginais da Binomial, Poisson, Binomial negativa, Gama e Normal inversa. Desse modo, são propostos os esquemas de perturbação aplicado diretamente no vetor de parâmetros que define a estrutura de dependência espacial, e na variável resposta a partir da média. Valida-

mos o critério proposto por um estudo de simulação. Os resultados mostraram que o critério proposto, em geral, é mais adequado para seleção da matriz de correlação de trabalho em comparação aos concorrentes.

Na Seção 4, propomos um modelo do tipo quasi-gama aumentada com zeros e espacialmente correlacionado. Assim, foi possível não só a modelagem do excesso de zeros presentes na amostra, mas também a dependência espacial. Em geral, se a amostra apresenta uma grande proporção de zeros, extrair informações sobre os dados torna-se uma tarefa relativamente difícil, sobretudo quando as respostas diferentes de zero são contínuas com uma certa assimetria, portanto, não gaussianas. Por conseguinte, o entendimento do comportamento dos dados torna-se importante quando a aplicação subjacente é, por exemplo, ecológica e/ou epidemiológica, como foi a motivação para tal. Com isso, desenvolvemos a modelagem para dados contínuos com assimetria positiva que possuem inflação de zeros e dependência espacial, através de uma quasi-verossimilhança aumentada com zeros capaz de acomodar a dependência espacial entre as observações. Assim, foram derivadas equações de estimação generalizadas e aplicadas no algoritmo Expectation-Solution (ES) para obter as estimativas dos parâmetros. Ainda, obtemos medidas de diagnóstico de influência local sob três esquemas de perturbação e sugerimos medidas para os resíduos. Finalmente, para confirmar a aplicabilidade da proposta, realizamos uma aplicação utilizando dados de precipitação pluviométrica acumulada no estado de Pernambuco, Brasil.

5.1 TRABALHOS FUTUROS

Muito ainda há de ser desenvolvido em análise de modelos geoestatísticos, principalmente no que diz respeito a análise de diagnóstico. No entanto, destaca-se alguns dos trabalhos futuros:

- i) Realizar estudos de simulação comparativos com outros cenários, para todas as propostas deste trabalho, a fim de avaliar o desempenho dos estimadores quanto à eficiência;
- ii) Desenvolver testes de hipóteses para os parâmetros dos modelo quasi-gama aumentada com zeros;
- iii) Sugerir um critério de seleção de modelos para auxiliar na escolha de covariáveis;

- iv) Verificar, a partir das medidas de influência local apresentadas na Seção 3, se a retirada de observações potencialmente influentes afetaria nos resultados dos critérios de seleção da matriz de correlação de trabalho;
- v) Utilizar dos resultados e rotinas computacionais desenvolvidas, para criar um pacote no software R, até então não disponível, para estimação dos parâmetros e interpolação com a suposição de normalidade e correlação Wendland. Ainda, um outro pacote para estimação por Equações de Estimação Generalizadas;
- vi) Pacote para modelagem com modelos geoestatísticos da quasi-gama aumenta com zero.

REFERÊNCIAS

- ALBERT, P. S.; MCSHANE, L. M. A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. *Biometrics*, JSTOR, p. 627–638, 1995.
- ANDERSON, T. et al. Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 1, n. 1, p. 135–141, 1973.
- ARTES, R. *Extensões da teoria das equações de estimação generalizadas a dados circulares e modelos de dispersão*. Tese (Doutorado) — Instituto de Matemática e Estatística da Universidade de São Paulo, 25/04/97., 1997.
- ASSUMPÇÃO, R. A. B.; OPAZO, M. A. U.; GALEA, M. Local influence for spatial analysis of soil physical properties and soybean yield using student's t-distribution. *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 35, p. 1917–1926, 2011.
- ATKINSON, A.; DONEV, A.; TOBIAS, R. *Optimum experimental designs, with SAS*. [S.l.]: Oxford University Press, 2007. v. 34.
- BACHOC, F.; PORCU, E.; BEVILACQUA, M.; FURRER, R.; FAOUZI, T. Asymptotically equivalent prediction in multivariate geostatistics. *arXiv preprint arXiv:2007.14684*, 2020.
- BANDEIRA, C. Z. R. et al. Modelando dados de contagem com inflação de zeros, sobredispersão e dependência espacial. Universidade Federal do Amazonas, 2015.
- BELSLEY, D. A.; KUH, E.; WELSCH, R. E. *Regression diagnostics: Identifying influential data and sources of collinearity*. [S.l.]: John Wiley & Sons, 2005. v. 571.
- BEVILACQUA, M.; DIGGLE, P. J.; PORCU, E. Families of covariance functions for bivariate random fields on spheres. *Spatial Statistics*, Elsevier, v. 40, p. 100448, 2020.
- BEVILACQUA, M.; FAOUZI, T.; FURRER, R.; PORCU, E. et al. Estimation and prediction using generalized wendland covariance functions under fixed domain asymptotics. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 47, n. 2, p. 828–856, 2019.
- BIONDI, F.; MYERS, D. E.; AVERY, C. C. Geostatistically modeling stem size and increment in an old-growth forest. *Canadian Journal of Forest Research*, NRC Research Press, v. 24, n. 7, p. 1354–1368, 1994.
- BONAT, W. H. Modelling mixed types of outcomes in additive genetic models. *The International Journal of Biostatistics*, De Gruyter, v. 13, n. 2, 2017.
- BORSSOI, J.; URIBE-OPAZO, M.; GALEA, M. Técnicas de diagnóstico de influência local na análise espacial da produtividade da soja. *Engenharia Agrícola*, v. 31, n. 2, p. 376–387, 2011.
- BORSSOI, J. A.; URIBE-OPAZO, M. A.; GALEA, M. Técnicas de diagnóstico de influência local na análise espacial da produtividade da soja. *Engenharia Agrícola*, SciELO Brasil, v. 31, n. 2, p. 376–387, 2011.

- BORSSOI, J. A.; URIBE-OPAZO, M. A.; ROJAS, M. G. Diagnostic techniques applied in geostatistics for agricultural data analysis. *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 33, p. 1561–1570, 2009.
- BUHMANN, M. D. *Radial basis functions: theory and implementations*. [S.l.]: Cambridge university press, 2003. v. 12.
- CANTONI, E.; FLEMMING, J. M.; RONCHETTI, E. Variable selection for marginal longitudinal generalized linear models. *Biometrics*, Wiley Online Library, v. 61, n. 2, p. 507–514, 2005.
- CARIO, M. C.; NELSON, B. L. *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix*. [S.l.], 1997.
- CASTRO, M. de; GALEA-ROJAS, M.; BOLFARINE, H.; CASTILHO, M. V. de. Local influence in regression models for the detection of analytical bias. *Chemometrics and intelligent laboratory systems*, Elsevier, v. 83, n. 2, p. 139–147, 2006.
- CEROVECKI, C.; FRANCQ, C.; HÖRMANN, S.; ZAKOIAN, J.-M. Functional garch models: the quasi-likelihood approach and its applications. *Journal of econometrics*, Elsevier, v. 209, n. 2, p. 353–375, 2019.
- CHANG, Y.-C. Residuals analysis of the generalized linear models for longitudinal data. *Statistics in Medicine*, Wiley Online Library, v. 19, n. 10, p. 1277–1293, 2000.
- CHANG, Y.-c. I. Strong consistency of maximum quasi-likelihood estimate in generalized linear models via a last time. *Statistics & Probability Letters*, Elsevier, v. 45, n. 3, p. 237–246, 1999.
- CHAPRA, S. C.; CANALE, R. P. *Métodos Numéricos para Engenharia-7ª Edição*. [S.l.]: McGraw Hill Brasil, 2016.
- CHERNIH, A.; HUBBERT, S. Closed form representations and properties of the generalised wendland functions. *Journal of Approximation Theory*, Elsevier, v. 177, p. 17–33, 2014.
- CHERNIH, A.; SLOAN, I. H.; WOMERSLEY, R. S. Wendland functions with increasing smoothness converge to a gaussian. *Advances in Computational Mathematics*, Springer, v. 40, n. 1, p. 185–200, 2014.
- CHU, B. Composite quasi-likelihood estimation of dynamic panels with group-specific heterogeneity and spatially dependent errors. *arXiv preprint arXiv:1704.06613*, 2017.
- COOK, D. O.; KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of proportions in finance with self selection. *Journal of empirical finance*, Elsevier, v. 15, n. 5, p. 860–867, 2008.
- COOK, R. D. Detection of influential observation in linear regression. *Technometrics*, Taylor & Francis Group, v. 19, n. 1, p. 15–18, 1977.
- COOK, R. D. Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 133–169, 1986.
- CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Piracicaba: USP*, 2013.

- CORDEIRO, G. M.; NETO, E. d. A. Modelos paramétricos. *Pernambuco: UFRPE*, 2004.
- CRESSIE, N. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, Springer, v. 17, n. 5, p. 563–586, 1985.
- CRESSIE, N. A. C. *Statistics for spatial data*. [S.l.]: Wiley series in probability and mathematical statistics. Applied probability and statistics section, 1993.
- DALEY, D. J.; PORCU, E.; BEVILACQUA, M. Classes of compactly supported covariance functions for multivariate random fields. *Stochastic Environmental Research and Risk Assessment*, Springer, v. 29, n. 4, p. 1249–1263, 2015.
- DALPOSSO, G. H.; URIBE-OPAZO, M. A.; BASTIANI, F. D. Spatial-temporal analysis of soybean productivity using geostatistical methods. *Journal of Agricultural Studies*, 2021.
- DAVISON, A. C.; HINKLEY, D. V. *Bootstrap methods and their application*. [S.l.]: Cambridge university press, 1997.
- DE BASTIANI, F. *Influência local em modelos espaciais lineares com distribuição da família de contornos elípticos*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2012.
- DE BASTIANI, F. Inference and diagnostics in spatial models. Universidade Federal de Pernambuco, 2016.
- DE BASTIANI, F.; CYSNEIROS, A. H. M. de A.; URIBE-OPAZO, M. A.; GALEA, M. Influence diagnostics in elliptical spatial linear models. *Test*, Springer, v. 24, n. 2, p. 322–340, 2015.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977.
- DENG, C.; GUAN, Y.; WAAGEPETERSEN, R. P.; ZHANG, J. Second-order quasi-likelihood for spatial point processes. *Biometrics*, Wiley Online Library, v. 73, n. 4, p. 1311–1320, 2017.
- DÍAZ-GARCÍA, J. A.; ROJAS, M. G.; LEIVA-SÁNCHEZ, V. Influence diagnostics for elliptical multivariate linear regression models. *Communications in statistics-Theory and Methods*, Taylor & Francis, v. 32, n. 3, p. 625–641, 2003.
- DIGGLE, P.; RIBEIRO, P. *Model based Geostatistics*. [S.l.]: Caxambu: Associação Brasileira de Estatística. 14 SINAPE - Simpósio Nacional de Probabilidade e Estatística, 2000.
- DIGGLE, P.; RIBEIRO, P. *Model-based Geostatistics*. [S.l.]: Springer series in statistics, 2007.
- DOBBIE MELISSA J E WELSH, A. H. Theory & methods: Modeling correlated zero-inflated count data. *Australian & New Zealand Journal of Statistics*, Wiley Online Library, v. 43, n. 4, p. 431–444, 2001.
- EFRON, B. Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 54, n. 1, p. 83–111, 1992.
- FASSHAUER, G. E. *Meshfree approximation methods with MATLAB*. [S.l.]: World Scientific, 2007. v. 6.

- FEUERVERGER, A. On some methods of analysis for weather experiments. *Biometrika*, Oxford University Press, v. 66, n. 3, p. 655–658, 1979.
- GALEA-ROJAS, M.; BOLFARINE, H.; CASTRO, M. de. Local influence in comparative calibration models. *Biometrical journal*, Wiley Online Library, v. 44, n. 1, p. 59–81, 2002.
- GENTON, M. G.; KLEIBER, W. Cross-covariance functions for multivariate geostatistics. *Statistical Science*, Institute of Mathematical Statistics, v. 30, n. 2, p. 147–163, 2015.
- GNEITING, T. Compactly supported correlation functions. *Journal of Multivariate Analysis*, Elsevier, v. 83, n. 2, p. 493–508, 2002.
- GNEITING, T. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, Taylor & Francis, v. 97, n. 458, p. 590–600, 2002.
- GNEITING, T. et al. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, Bernoulli Society for Mathematical Statistics and Probability, v. 19, n. 4, p. 1327–1349, 2013.
- GODAMBE, V. P.; HEYDE, C. C. Quasi-likelihood and optimal estimation. In: *Selected works of cc heyde*. [S.l.]: Springer, 2010. p. 386–399.
- GOSHO, M.; HAMADA, C.; YOSHIMURA, I. Criterion for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 40, n. 21, p. 3839–3856, 2011.
- GUAN, Y.; JALILIAN, A.; WAAGEPETERSEN, R. Quasi-likelihood for spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 77, n. 3, p. 677–697, 2015.
- HALL, D. B.; ZHANG, Z. Marginal models for zero inflated clustered data. *Statistical Modelling*, Sage Publications Sage CA: Thousand Oaks, CA, v. 4, n. 3, p. 161–180, 2004.
- HELLER, G.; STASINOPOULOS, M.; RIGBY, B. et al. The zero-adjusted inverse gaussian distribution as a model for insurance claims. In: GALWAY. *Proceedings of the 21th International Workshop on Statistical Modelling*. [S.l.], 2006. v. 226233.
- HIN, L.-Y.; CAREY, V. J.; WANG, Y.-G. Criteria for working–correlation–structure selection in gee: Assessment via simulation. *The American Statistician*, Taylor & Francis, v. 61, n. 4, p. 360–364, 2007.
- HIN, L.-Y.; WANG, Y.-G. Working-correlation-structure identification in generalized estimating equations. *Statistics in medicine*, Wiley Online Library, v. 28, n. 4, p. 642–658, 2009.
- HUZII, M. Some properties of conditional quasi-likelihood functions for time series model fitting. *Journal of Time Series Analysis*, Wiley Online Library, v. 9, n. 4, p. 345–353, 1988.
- IBGE. Instituto brasileiro de geografia e estatística. cidade e estados. Disponível em: <<https://www.ibge.gov.br/cidades-e-estados/ba.html>> Acesso em 17/05/2021., 2020.
- INAN, G.; LATIF, M. A.; PREISSER, J. A prediction criterion for working correlation structure selection in gee. *arXiv preprint arXiv:1803.06383*, 2018.

- JAMAN, A.; LATIF, M. A.; BARI, W.; WAHED, A. S. A determinant-based criterion for working correlation structure selection in generalized estimating equations. *Statistics in medicine*, Wiley Online Library, v. 35, n. 11, p. 1819–1833, 2016.
- JEONG, J.; JUN, M. A class of matérn-like covariance functions for smooth processes on a sphere. *Spatial Statistics*, Elsevier, v. 11, p. 1–18, 2015.
- JUNG, K. M. Local influence of the quasi-likelihood estimators in generalized linear models. *Communications for Statistical Applications and Methods*, The Korean Statistical Society, and Korean International Statistical Society, v. 14, n. 1, p. 229–239, 2007.
- JUNG, K.-m. Local influence in generalized estimating equations. *Scandinavian journal of statistics*, Wiley Online Library, v. 35, n. 2, p. 286–294, 2008.
- LI, S.; CAI, T. T.; LI, H. Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *Journal of the American Statistical Association*, Taylor & Francis, p. 1–33, 2021.
- LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, Oxford University Press, v. 73, n. 1, p. 13–22, 1986.
- LIANG, M.; MARCOTTE, D. A class of non-stationary covariance functions with compact support. *Stochastic environmental research and risk assessment*, Springer, v. 30, n. 3, p. 973–987, 2016.
- LIN, P.-S.; CLAYTON, M. K. et al. Analysis of binary spatial data by quasi-likelihood estimating equations. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 33, n. 2, p. 542–555, 2005.
- MANUEL, L.; SCALON, J. D. Generalized estimating equations approach for spatial lattice data: A case study in adoption of improved maize varieties in mozambique. *Biometrical Journal*, Wiley Online Library, v. 62, n. 8, p. 1879–1895, 2020.
- MARTIN, M. A.; ROBERTS, S. Jackknife-after-bootstrap regression influence diagnostics. *Journal of Nonparametric Statistics*, Taylor & Francis, v. 22, n. 2, p. 257–269, 2010.
- MARTINEZ, R. O. *Modelos de regressão beta inflacionados*. Tese (Doutorado) — Universidade de São Paulo, 2008.
- MATÉRN, B. Spatial variation: Stochastic previous models and their application to some problems in forest surveys and other sampling investigations. *Medd. Statens Skogsforskningsinstitut*, v. 49, n. 5, 1960.
- MATHERON, G. Les variables régionalisées et leur estimation [regionalized variables and their estimation]. *Paris: Editions Masson*, 1965.
- MCCULLAGH, P. Quasi-likelihood functions. *The Annals of Statistics*, JSTOR, p. 59–67, 1983.
- MCCULLAGH, P.; NELDER, J. Generalized linear models., 2nd edn.(chapman and hall: London). *Standard book on generalized linear models*, 1989.
- MONOD, A. Generalized estimating equations for zero-inflated spatial count data. *Procedia Environmental Sciences*, Elsevier, v. 7, p. 281–286, 2011.

- MOREAUX, G. Compactly supported radial covariance functions. *Journal of Geodesy*, Springer, v. 82, n. 7, p. 431–443, 2008.
- MYERS, D.; IACO, S. D.; POSA, D.; CESARE, L. D. Space-time radial basis functions. *Computers & Mathematics with Applications*, Elsevier, v. 43, n. 3-5, p. 539–549, 2002.
- NAVA, D.; BASTIANI, F. D.; URIBE-OPAZO, M.; NICOLIS, O.; GALEA, M. Local influence for spatially correlated binomial data: An application to the spodoptera frugiperda infestation in corn. *Journal of Agricultural, Biological and Environmental Statistics*, Springer, v. 22, n. 4, p. 540–561, 2017.
- NELDER, J. A.; PREGIBON, D. An extended quasi-likelihood function. *Biometrika*, Oxford University Press, v. 74, n. 2, p. 221–232, 1987.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society Series A*, v. 135, p. 370384, 1972.
- NIAKI, S. T. A.; ABBASI, B. Generating correlation matrices for normal random vectors in norta algorithm using artificial neural networks. *Journal of Uncertain Systems*, Citeseer, v. 2, n. 3, p. 192–201, 2008.
- ORTEGA, E. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. *Computational statistics & data analysis*, Elsevier, v. 42, n. 1-2, p. 165–186, 2003.
- OSPINA, R.; FERRARI, S. L. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 6, p. 1609–1623, 2012.
- PAN, W. Akaike's information criterion in generalized estimating equations. *Biometrics*, Wiley Online Library, v. 57, n. 1, p. 120–125, 2001.
- PAULA, G. A. Assessing local influence in restricted regression models. *Computational statistics & data analysis*, Elsevier, v. 16, n. 1, p. 63–79, 1993.
- PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2013.
- POON, W.-Y.; POON, Y. S. Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 61, n. 1, p. 51–61, 1999.
- PORCU, E.; BEVILACQUA, M.; GENTON, M. G. Nonseparable, space-time covariance functions with dynamical compact supports. Academia Sinica, Institute of Statistical Science, 2020.
- PORCU, E.; DALEY, D. J.; BUHMANN, M.; BEVILACQUA, M. Radial basis functions with compact support for multivariate geostatistics. *Stochastic environmental research and risk assessment*, Springer, v. 27, n. 4, p. 909–922, 2013.
- PORCU, E.; ZASTAVNYI, V. P.; BEVILACQUA, M. Buhmann covariance functions, their compact supports, and their smoothness. *Dolomites Research Notes on Approximation*, v. 10, n. 1, 2017.

- PRENTICE, R. L. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, JSTOR, p. 1033–1048, 1988.
- PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. *Numerical recipes 3rd edition: The art of scientific computing*. [S.l.]: Cambridge university press, 2007.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <<https://www.R-project.org/>>.
- RODRIGUES-MOTTA, M.; SOTO, D. M. G.; LACHOS, V. H.; VILCA, F.; BALTAR, V. T.; JUNIOR, E. V.; FISBERG, R. M.; MARCHIONI, D. M. L. A mixed-effect model for positive responses augmented by zeros. *Statistics in medicine*, Wiley Online Library, v. 34, n. 10, p. 1761–1778, 2015.
- ROSEN, O.; JIANG, W.; TANNER, M. A. Mixtures of marginal models. *Biometrika*, Oxford University Press, v. 87, n. 2, p. 391–404, 2000.
- ROTNITZKY, A.; JEWELL, N. P. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, Oxford University Press, v. 77, n. 3, p. 485–497, 1990.
- SANTOS, M. dos. Estimação por máxima verossimilhança para manter o padrão dos parâmetros da distribuição weibull via bfgs com a linguagem de programação ox. *REMAT: Revista Eletrônica da Matemática*, v. 5, n. 1, p. 84–100, 2019.
- SATO, T.; MATSUDA, Y. Spatial extension of generalized autoregressive conditional heteroskedasticity models. *Spatial Economic Analysis*, Taylor & Francis, p. 1–13, 2020.
- SCHABACK, R. The missing wendland functions. *Advances in Computational Mathematics*, Springer, v. 34, n. 1, p. 67–81, 2011.
- SEIDEL, E. J.; OLIVEIRA, M. S. d. Novo índice geoestatístico para a mensuração da dependência espacial. *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 38, n. 3, p. 699–705, 2014.
- SEIDEL, E. J.; OLIVEIRA, M. S. de. Medidas de dependência espacial baseadas em duas perspectivas do semivariograma paramétrico. *Ciência e Natura*, Universidade Federal de Santa Maria, v. 37, n. 3, p. 20–27, 2015.
- SHULTS, J.; CHAGANTY, N. R. Analysis of serially correlated data using quasi-least squares. *Biometrics*, JSTOR, p. 1622–1630, 1998.
- SILVA, J. A. D.; CIRILLO, M. A. Selection criterion of work matrix as a function of limiting estimates of the covariance matrix of correlated data in gee. *Biometrical Journal*, Wiley Online Library, v. 60, n. 5, p. 979–990, 2018.
- STASINOPOULOS, D. M.; RIGBY, R. A. et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007.
- STASINOPOULOS, M.; RIGBY, B.; AKANTZILIOTOU, C. *Instructions on how to use the gamlss package in R Second Edition*. 2008.
- STEIN, M. L. *Interpolation of spatial data: some theory for kriging*. [S.l.]: NY: Springer, 1999.

- STIRATELLI, R.; LAIRD, N.; WARE, J. H. Random-effects models for serial observations with binary response. *Biometrics*, JSTOR, p. 961–971, 1984.
- SUTRADHAR, B. C.; DAS, K. On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika*, Oxford University Press, v. 86, n. 2, p. 459–465, 1999.
- SUTRADHAR, B. C.; DAS, K. On the accuracy of efficiency of estimating equation approach. *Biometrics*, Wiley Online Library, v. 56, n. 2, p. 622–625, 2000.
- SUTRADHAR, B. C.; OYET, A. J. An unbalanced multidimensional latent effects-based logistic mixed model and ggl estimation for spatial binary data. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 90, n. 18, p. 3338–3359, 2020.
- URIBE-OPAZO, M. A.; BORSSOI, J. A.; GALEA, M. Influence diagnostics in gaussian spatial linear models. *Journal of Applied Statistics*, Taylor & Francis, v. 39, n. 3, p. 615–630, 2012.
- VENEZUELA, M. K.; ARTES, R. Estimating equations and diagnostic techniques applied to zero-inflated models for panel data. *Electronic Journal of Statistics*, Institute of Mathematical Statistics and Bernoulli Society, v. 8, n. 1, p. 1641–1660, 2014.
- VENEZUELA, M. K.; BOTTER, D. A.; SANDOVAL, M. C. Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 77, n. 10, p. 879–888, 2007.
- VENEZUELA, M. K.; SANDOVAL, M. C.; BOTTER, D. A. Local influence in estimating equations. *Computational statistics & data analysis*, Elsevier, v. 55, n. 4, p. 1867–1883, 2011.
- VOS, P. W. Minimum f-divergence estimators and quasi-likelihood functions. *Annals of the Institute of Statistical Mathematics*, Springer, v. 44, n. 2, p. 261–279, 1992.
- WEDDERBURN, R. W. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, Oxford University Press, v. 61, n. 3, p. 439–447, 1974.
- WENDLAND, H. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, Springer, v. 4, n. 1, p. 389–396, 1995.
- WENDLAND, H. *Scattered Data Approximation*. Cambridge University Press, 2004. (Cambridge Monographs on Applied and Computational Mathematics). ISBN 0521843359,9780521843355,9780511265792. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=5D86B68EC8200B80DE10809C2698DC03>.
- WONG, W. K. Comparing robust properties of a, d, e and g-optimal designs. *Computational statistics & data analysis*, Elsevier, v. 18, n. 4, p. 441–448, 1994.
- WU, Z. Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, Springer, v. 4, n. 1, p. 283, 1995.
- YAHAV, I.; SHMUELI, G. On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, Wiley Online Library, v. 28, n. 1, p. 91–102, 2012.

ZASTAVNYI, V. On some properties of buhmann functions. *Ukrainian Mathematical Journal*, Springer, v. 58, n. 8, p. 1184–1208, 2006.

ZEGER, S. L.; LIANG, K.-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, JSTOR, p. 121–130, 1986.

ZEGER, S. L.; LIANG, K.-Y.; ALBERT, P. S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, JSTOR, p. 1049–1060, 1988.

ZEGER, S. L.; LIANG, K.-Y.; SELF, S. G. The analysis of binary longitudinal data with time independent covariates. *Biometrika*, Oxford University Press, v. 72, n. 1, p. 31–38, 1985.

ZHANG, H. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, Taylor & Francis, v. 99, n. 465, p. 250–261, 2004.

ZHANG, X.; KING, M. L. Influence diagnostics in generalized autoregressive conditional heteroscedasticity processes. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 23, n. 1, p. 118–129, 2005.

ZHAO, L. P.; PRENTICE, R. L.; SELF, S. G. Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 54, n. 3, p. 805–811, 1992.

ZHU, H.; IBRAHIM, J. G.; LEE, S.; ZHANG, H. et al. Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 35, n. 6, p. 2565–2588, 2007.

ZHU, H.; ZHANG, H. A diagnostic procedure based on local influence. *Biometrika*, Oxford University Press, v. 91, n. 3, p. 579–589, 2004.

ZHU, H.-T.; LEE, S.-Y. Local influence for incomplete data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 63, n. 1, p. 111–126, 2001.

Analyzing longitudinal claims cost data: A zero-augmented gamma mixed regression approach. v. 15.

APÊNDICE A – RESULTADOS DE ML

Neste Apêndice estão os seguintes resultados: Logaritmo da função de verossimilhança para um modelo espacial linear Gaussiano e correlação de Wendland e suas derivadas com relação aos parâmetros; detalhes dos cálculos para se obter a matriz de informação de Fisher.

$$\begin{aligned}
 \log L(\boldsymbol{\beta}, \boldsymbol{\phi}) &= l(\boldsymbol{\beta}, \boldsymbol{\phi}) = \log \left\{ \frac{1}{|\boldsymbol{\Sigma}|^{1/2} \sqrt{(2\pi)^n}} \exp \left[-\frac{(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})}{2} \right] \right\} \\
 &= \log (|\boldsymbol{\Sigma}|^{-1/2}) + \log ((2\pi)^{-n/2}) + \log \left\{ \exp \left[-\frac{(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{X} - \mathbf{X}\boldsymbol{\beta})}{2\boldsymbol{\Sigma}} \right] \right\} \\
 &= -\frac{1}{2} \log (|\boldsymbol{\Sigma}|) - \frac{n}{2} \log(2\pi) - \frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}),
 \end{aligned}$$

Derivada com relação a $\boldsymbol{\beta}$.

$$\begin{aligned}
 \frac{\partial l}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{\Sigma}^{-1} \mathbf{Z} - \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta}) \right\} \\
 &= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2} [\mathbf{Z}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta}] \right\} \\
 &= -\frac{1}{2} (2\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z}) - \frac{1}{2} 2\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta} \\
 &= \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z} - \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta},
 \end{aligned}$$

Derivadas com relação ao parâmetro $\boldsymbol{\phi}$.

$$\begin{aligned}
 \frac{\partial l}{\partial \phi_i} &= -\frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \right) - \frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \phi_i} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \\
 &= -\frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \right) + \frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}),
 \end{aligned}$$

Derivadas de segunda ordem.

$$\begin{aligned}
 \frac{\partial^2 l}{\partial \phi_i \partial \phi_j} &= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \right) + \frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \mathbf{Z} - \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta} \right) \right\} \\
 &= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \frac{1}{2} \mathbf{Z}^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \mathbf{Z} - \boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \frac{1}{2} \boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta} \right\} \\
 &= \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \mathbf{Z}
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial^2 l}{\partial \phi_i \partial \phi_j} &= -\frac{1}{2} \text{tr} \left(-\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_j} + \boldsymbol{\Sigma}^{-1} \frac{\partial^2 \boldsymbol{\Sigma}}{\partial \phi_i \partial \phi_j} \right) + \\
 &\quad \frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \left[-\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_j} \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \frac{\partial^2 \boldsymbol{\Sigma}}{\partial \phi_i \partial \phi_j} \boldsymbol{\Sigma}^{-1} - \right. \\
 &\quad \left. \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_j} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i} \boldsymbol{\Sigma}^{-1} \right] (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})
 \end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{2}tr \left(-\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} + \Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} \right) + \\
&\quad \frac{1}{2}(\mathbf{Z} - \mathbf{X}\beta)^\top \Sigma^{-1} \left[-\frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} + \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} - \frac{\partial \Sigma}{\partial \phi_j} \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \right] \Sigma^{-1}(\mathbf{Z} - \mathbf{X}\beta).
\end{aligned}$$

Valor esperado das derivadas de segunda ordem.

$$\begin{aligned}
E \left(-\frac{\partial^2 l}{\partial \phi_j \partial \phi_i} \right) &= \frac{1}{2} E \left[tr \left(\Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} \right) \right] \\
&\quad - \frac{1}{2} E \left[tr \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \right) \right] \\
&= \frac{1}{2} tr \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \right),
\end{aligned}$$

e

$$\begin{aligned}
E \left(-\frac{\partial^2 l}{\partial \beta \partial \phi_i} \right) &= E \left(-\frac{\partial^2 l}{\partial \phi_i \partial \beta} \right) \\
&= E \left[- \left(\mathbf{X}^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \mathbf{X} \beta - \mathbf{X}^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \mathbf{Z} \right) \right] \\
&= -\mathbf{X}^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \mathbf{X} \beta + \mathbf{X}^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} E(\mathbf{Z}) \\
&= -\mathbf{X}^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \mathbf{X} \beta + \mathbf{X}^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \mathbf{X} \beta \\
&= \mathbf{0}
\end{aligned}$$

APÊNDICE B – SISTEMAS DE EQUAÇÕES PARA O MÉTODO DOS MÍNIMOS QUADRADOS

Neste Apêndice estão presentes os resultados das expressões para os sistemas de equações, para as estimativas dos parâmetros de uma semivariograma por mínimos quadrados ordinários.

$$\frac{\partial Q(\mathbf{h}, \boldsymbol{\phi})_0}{\partial \phi_i} = \begin{cases} -2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^2 \right) \right] = 0 \\ -2 \sum_{i,j=1}^m \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^2 \right) \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^2 \right) \right] = 0 \\ 4 \sum_{i,j=1}^m \frac{\phi_2 h_{ij}}{\phi_3^2} \left(1 - \frac{h_{ij}}{\phi_3} \right) \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^2 \right) \right] = 0 \end{cases} \quad (\text{B.1})$$

$$\frac{\partial Q(\mathbf{h}, \boldsymbol{\phi})_1}{\partial \phi_i} = \begin{cases} -2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^4 \left(4 \frac{h_{ij}}{\phi_3} + 1 \right) \right) \right] = 0 \\ -2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^4 \left(4 \frac{h_{ij}}{\phi_3} + 1 \right) \right) \right] \times A = 0 \\ 2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^4 \left(4 \frac{h_{ij}}{\phi_3} + 1 \right) \right) \right] \times B = 0 \end{cases} \quad (\text{B.2})$$

em que $A = \left(1 - \left(1 - \frac{h_{ij}}{\phi_3} \right)^4 \left(4 \frac{h_{ij}}{\phi_3} + 1 \right) \right)$

e $B = \left[4 \frac{\phi_2}{\phi_3^2} \left(1 - \frac{h_{ij}}{\phi_3} \right)^3 \left(4 \frac{h_{ij}}{\phi_3} + 1 \right) - 4 \phi_2 \frac{h_{ij}}{\phi_3^2} \left(1 - \frac{h_{ij}}{\phi_3} \right)^4 \right]$.

$$\frac{\partial Q(\mathbf{h}, \boldsymbol{\phi})_2}{\partial \phi_i} = \begin{cases} -2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - f \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^6 \right) \right] = 0 \\ -2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - f \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^6 \right) \right] \times C = 0 \\ 2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - f \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^6 \right) \right] \times D = 0, \end{cases} \quad (\text{B.3})$$

em que $f = \left(1 + \frac{6h_{ij}}{\phi_3} + \frac{35h_{ij}^2}{\phi_3^2} \right)$, $C = \left(1 - f \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^6 \right)$ e

$$D = \left[\phi_2 \left(-\frac{6h_{ij}}{\phi_3^2} - \frac{70h_{ij}^2}{\phi_3^3} \right) \left(1 - \frac{h_{ij}}{\phi_3} \right)^6 + \frac{6\phi_2 h_{ij}}{\phi_3^2} \left(1 + \frac{6h_{ij}}{\phi_3} + \frac{35h_{ij}^2}{\phi_3^2} \right) \left(1 - \frac{h_{ij}}{\phi_3} \right)^5 \right].$$

$$\frac{\partial Q(\mathbf{h}, \phi)_3}{\partial \phi_i} = \begin{cases} -2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - g \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^8 \right) \right] = 0 \\ -2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - g \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^8 \right) \right] \times E = 0 \\ 2 \sum_{i,j=1}^m \left[\gamma(h_{ij}) - \phi_1 - \phi_2 \left(1 - g \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^8 \right) \right] \times G = 0, \end{cases} \quad (\text{B.4})$$

em que $g = \left(1 + \frac{8h_{ij}}{\phi_3} + \frac{25h_{ij}^2}{\phi_3^2} + \frac{32h_{ij}^3}{\phi_3^3} \right)$, $E = \left[1 - g \times \left(1 - \frac{h_{ij}}{\phi_3} \right)^8 \right]$ e,

$$G = \left[\phi_2 \left(-\frac{8h_{ij}}{\phi_3^2} - \frac{50h_{ij}^2}{\phi_3^3} - \frac{96h_{ij}^3}{\phi_3^4} \right) \left(1 - \frac{h_{ij}}{\phi_3} \right)^8 + \frac{8g\phi_2 h_{ij}}{\phi_3^2} \left(1 - \frac{h_{ij}}{\phi_3} \right)^7 \right].$$

APÊNDICE C – DERIVADAS DA FUNÇÃO DE COVARIÂNCIA DE WENDLAND

Considere a função de covariância pertencente a família de Wendland, $C(h)$, com $h = \|s_i - s_j\|$, definida a partir da Seção 2.3.1 para os parâmetros de suavização $\kappa = 0, 1, 2$ e 3 . Ainda, considere o vetor de parâmetros que define a estrutura espacial, $\phi = (\phi_1, \phi_2, \phi_3)$, e o parâmetro $\mu = \lceil \frac{3}{2} + \kappa \rceil$, definido a partir da relação (2.3) em que $\lceil \cdot \rceil$ denota a função teto. Então,

i - Para $\kappa = 0 \Rightarrow \mu = 2$, assim,

$$C(h) = \phi_2 \left(1 - \frac{h}{\phi_3}\right)^2,$$

com

$$\frac{\partial C}{\partial \phi_1} = 0; \quad \frac{\partial C}{\partial \phi_2} = \left(1 - \frac{h}{\phi_3}\right)^2;$$

$$\frac{\partial C}{\partial \phi_3} = 2\phi_2 \left(1 - \frac{h}{\phi_3}\right) \frac{h}{\phi_3^2};$$

ii - Para $\kappa = 1 \Rightarrow \mu = 3$, assim,

$$C(h) = \phi_2 \left(4\frac{h}{\phi_3} + 1\right) \left(1 - \frac{h}{\phi_3}\right)^4,$$

com

$$\frac{\partial C}{\partial \phi_1} = 0; \quad \frac{\partial C}{\partial \phi_2} = \left(4\frac{h}{\phi_3} + 1\right) \left(1 - \frac{h}{\phi_3}\right)^4;$$

$$\frac{\partial C}{\partial \phi_3} = 4\phi_2 \frac{h}{\phi_3^2} \left(4\frac{h}{\phi_3} + 1\right) \left(1 - \frac{h}{\phi_3}\right)^3 + \phi_2 \left(1 - \frac{h}{\phi_3}\right)^4 \left(-4\frac{h}{\phi_3^2}\right);$$

iii - Para $\kappa = 2 \Rightarrow \mu = 4$, assim,

$$C(h) = \phi_2 \left(1 + 6\frac{h}{\phi_3} + 35\frac{h^2}{3\phi_3^2}\right) \left(1 - \frac{h}{\phi_3}\right)^6,$$

com

$$\frac{\partial C}{\partial \phi_1} = 0; \quad \frac{\partial C}{\partial \phi_2} = \left(1 + 6\frac{h}{\phi_3} + 35\frac{h^2}{3\phi_3^2}\right) \left(1 - \frac{h}{\phi_3}\right)^6;$$

$$\frac{\partial C}{\partial \phi_3} = 6\phi_2 \frac{h}{\phi_3^2} \left(1 + 6\frac{h}{\phi_3} + 35\frac{h^2}{3\phi_3^2}\right) \left(1 - \frac{h}{\phi_3}\right)^5 + \phi_2 \left(1 - \frac{h}{\phi_3}\right)^6 \left(-6\frac{h}{\phi_3^2} - 70\frac{h^2}{9\phi_3^3}\right);$$

iv - Para $\kappa = 3 \Rightarrow \mu = 5$, assim,

$$C(h) = \phi_2 \left(1 + 8 \frac{h}{\phi_3} + 25 \frac{h^2}{\phi_3^2} + 32 \frac{h^3}{\phi_3^3} \right) \left(1 - \frac{h}{\phi_3} \right)^8,$$

com

$$\frac{\partial C}{\partial \phi_1} = 0; \quad \frac{\partial C}{\partial \phi_2} = \left(1 + 8 \frac{h}{\phi_3} + 25 \frac{h^2}{\phi_3^2} + 32 \frac{h^3}{\phi_3^3} \right) \left(1 - \frac{h}{\phi_3} \right)^8;$$

$$\begin{aligned} \frac{\partial C}{\partial \phi_3} = & 8\phi_2 \frac{h}{\phi_3^2} \left(1 + 8 \frac{h}{\phi_3} + 25 \frac{h^2}{\phi_3^2} + 32 \frac{h^3}{\phi_3^3} \right) \left(1 - \frac{h}{\phi_3} \right)^7 + \\ & \phi_2 \left(-8 \frac{h}{\phi_3^2} - 50 \frac{h^2}{\phi_3^3} - 96 \frac{h^3}{\phi_3^4} \right) \left(1 - \frac{h}{\phi_3} \right)^8. \end{aligned}$$

APÊNDICE D – DETALHES DOS RESULTADOS APRESENTADOS NA SEÇÃO 3.5 PARA MEDIDAS DE INFLUÊNCIA LOCAL

Neste apêndice são apresentados os detalhes dos resultados listados na Tabela 16 referente as expressões para o cálculo das medidas de influência local.

I. Binomial: $B(m, \mu)$. Função de ligação $g(\mu) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{X}\beta$, em que $\pi = (\pi_1, \dots, \pi_n)$ é um vetor das probabilidades de sucesso. Então,

$$\frac{\pi}{1-\pi} = \exp(\eta) \Rightarrow \pi = \frac{\exp(\eta)}{1 + \exp(\eta)} \Rightarrow 1 - \pi = \frac{1}{1 + \exp(\eta)} \quad \text{então} \quad \mu = m \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

$$\frac{\partial \mu}{\partial \beta_j} = \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta_j} = m\pi(1 - \pi)x_j$$

$$\Rightarrow \dot{\mathbf{K}} = \text{diag} [m_i\pi_i(1 - \pi_i)] \Rightarrow \boxed{\frac{\partial \dot{\mathbf{K}}}{\partial \beta_j} = \text{diag} [m_i\pi_i(1 - \pi_i)(1 - 2\pi_i)x_{ij}]}.$$

função de variância $\mathbf{v}_\mu = \text{diag} [m_i\pi_i(1 - \pi_i)]$ então,

$$\frac{\partial \mathbf{v}_\mu^{1/2}}{\partial \beta_j} = \text{diag} \left[\frac{1}{2} (m_i\pi_i(1 - \pi_i))^{-1/2} \frac{\partial v_{\mu_i}}{\partial \eta_i} x_{ij} \right].$$

II. Poisson: $P(\mu)$. Função de ligação $g(\mu) = \log(\mu) \Rightarrow \mu = \exp(\eta)$.

$$\frac{\partial \mu}{\partial \beta_j} = \exp(\eta)x_j \Rightarrow \dot{\mathbf{K}} = \text{diag} [\exp(\eta_i)] \Rightarrow \frac{\partial \dot{\mathbf{K}}}{\partial \beta_j} = \text{diag} [\exp(\eta_i)x_{ij}] \quad ,$$

função de variância $\mathbf{v}_\mu = \text{diag} [\mu_i]$ então, $\frac{\partial \mathbf{v}_\mu^{1/2}}{\partial \beta_j} = \text{diag} \left[\frac{1}{2} \exp(\eta_i)^{1/2} x_{ij} \right].$

III. Binomial Negativa: $BN(\mu, m)$. Função de ligação $g(\mu) = \log(\pi) \Rightarrow \mu = m \exp(\eta)$ com a probabilidade de sucesso em cada realização sendo $\pi_i = \exp(\eta_i)$.

$$\begin{aligned} \frac{\partial \mu}{\partial \beta_j} &= m\pi x_j \Rightarrow \dot{\mathbf{K}} = \text{diag} [m_i\pi_i] \\ &\Rightarrow \frac{\partial \dot{\mathbf{K}}}{\partial \beta_j} = \text{diag} [m_i\pi_i x_{ij}], \end{aligned}$$

função de variância $\mathbf{v}_\mu = \text{diag} [m_i\pi_i(1 + \pi_i)]$ então,

$$\frac{\partial \mathbf{v}_\mu^{1/2}}{\partial \beta_j} = \text{diag} \left[m_i \frac{1}{2} (m_i\pi_i(1 + \pi_i))^{-1/2} (m_i\pi_i(1 + 2\pi_i))x_{ij} \right].$$

IV. Gama: $G(\mu, \nu)$, em que ν é um parâmetro de forma e $\mu = \frac{\nu}{\phi}$, com ϕ um parâmetro de escala. Função de ligação $g(\mu) = \frac{1}{\mu}$.

$$\frac{\partial \mu}{\partial \beta_j} = -\frac{1}{\eta^2} x_j \Rightarrow \dot{\mathbf{K}} = \text{diag} \left[-\frac{1}{\eta_i^2} \right]$$

$$\Rightarrow \frac{\partial \dot{\mathbf{K}}}{\partial \beta_j} = \text{diag} \left[\frac{2}{\eta_i^3} x_{ij} \right],$$

função de variância $\mathbf{v}_{\boldsymbol{\mu}} = \text{diag} [\mu_i^2]$ então,

$$\frac{\partial \mathbf{v}_{\boldsymbol{\mu}}^{1/2}}{\partial \beta_j} = \text{diag} \left[-\frac{1}{\eta_i^2} \mathbf{x}_{ij} \right].$$

V. Normal Inversa: $NI(\boldsymbol{\mu})$. Função de ligação $g(\boldsymbol{\mu}) = \frac{1}{\boldsymbol{\mu}^2}$.

$$\frac{\partial \boldsymbol{\mu}}{\partial \beta_j} = -\frac{1}{2} \boldsymbol{\eta}^{-5/2} \mathbf{x}_j \Rightarrow \dot{\mathbf{K}} = \text{diag} \left[-\frac{1}{2} \eta_i^{-5/2} \right]$$

$$\Rightarrow \frac{\partial \dot{\mathbf{K}}}{\partial \beta_j} = \text{diag} \left[\frac{5}{4} \eta_i^{-7/2} x_{ij} \right],$$

função de variância $\mathbf{v}_{\boldsymbol{\mu}} = \text{diag} [\mu_i^3]$ então,

$$\frac{\partial \mathbf{v}_{\boldsymbol{\mu}}^{1/2}}{\partial \beta_j} = \text{diag} \left[-\frac{3}{2} \eta_i^{-5/2} \mathbf{x}_{ij} \right].$$

**APÊNDICE E – DERIVADAS PARA FUNÇÕES DE COVARIÂNCIA DE
WENDLAND, COM RELAÇÃO A PERTURBAÇÃO ω_i , PARA VÁRIOS VALORES
DE κ E μ**

O parâmetro μ é estabelecido de tal modo que $\mu = \lceil \frac{3}{2} + \kappa \rceil$, definido a partir da relação (2.3). Então, algumas funções para $\rho_{\mu,\kappa}(\frac{\omega_i h}{\phi_3})$ são dadas por:

i - Para $\kappa = 0 \Rightarrow \mu = 2$, tem-se $\rho_{2,0}(\frac{\omega_i h}{\phi_3}) = \left(1 - \frac{h\omega_i}{\phi_3}\right)^2$,

ii - Para $\kappa = 1 \Rightarrow \mu = 3$, tem-se

$$\rho_{3,1}(\frac{\omega_i h}{\phi_3}) = \left(4\frac{h\omega_i}{\phi_3} + 1\right) \left(1 - \frac{h\omega_i}{\phi_3}\right)^4,$$

iii - Para $\kappa = 2 \Rightarrow \mu = 4$, tem-se,

$$\rho_{4,2}(\frac{\omega_i h}{\phi_3}) = \left(1 + 6\frac{h\omega_i}{\phi_3} + 35\frac{h^2\omega_i^2}{\phi_3^2}\right) \left(1 - \frac{h\omega_i}{\phi_3}\right)^6,$$

iv - Para $\kappa = 3 \Rightarrow \mu = 5$, tem-se,

$$\rho_{5,3}(\frac{\omega_i h}{\phi_3}) = \left(1 + 8\frac{h\omega_i}{\phi_3} + 25\frac{h^2\omega_i^2}{\phi_3^2} + 32\frac{h^3\omega_i^3}{\phi_3^3}\right) \left(1 - \frac{h\omega_i}{\phi_3}\right)^8.$$

Portanto, suas respectivas derivadas com relação ao parâmetro de perturbação são:

$$\frac{\partial \rho_{2,0}(\frac{\omega_i h}{\phi_3})}{\partial \omega_i} = -2\frac{h}{\phi_3} \left(1 - \frac{h\omega_i}{\phi_3}\right),$$

$$\frac{\partial \rho_{3,1}(\frac{\omega_i h}{\phi_3})}{\partial \omega_i} = 4\frac{h}{\phi_3} \left[\left(1 - \frac{h\omega_i}{\phi_3}\right)^4 - \left(1 - \frac{h\omega_i}{\phi_3}\right)^3 \left(4\frac{h\omega_i}{\phi_3} + 1\right) \right],$$

$$\begin{aligned} \frac{\partial \rho_{4,2}(\frac{\omega_i h}{\phi_3})}{\partial \omega_i} &= -6\frac{h}{\phi_3} \left(1 - \frac{h\omega_i}{\phi_3}\right)^5 \left(1 + 6\frac{h\omega_i}{\phi_3} + 35\frac{h^2\omega_i^2}{\phi_3^2}\right) \\ &\quad + \left(1 - \frac{h\omega_i}{\phi_3}\right)^6 \left(6\frac{h}{\phi_3} + 70\frac{h^2\omega_i}{\phi_3^2}\right), \end{aligned}$$

$$\begin{aligned} \frac{\partial \rho_{5,3}(\frac{\omega_i h}{\phi_3})}{\partial \omega_i} &= -8\frac{h}{\phi_3} \left(1 - \frac{h\omega_i}{\phi_3}\right)^7 \left(1 + 8\frac{h\omega_i}{\phi_3} + 25\frac{h^2\omega_i^2}{\phi_3^2} + 32\frac{h^3\omega_i^3}{\phi_3^3}\right) \\ &\quad + \left(1 - \frac{h\omega_i}{\phi_3}\right)^8 \left(8\frac{h}{\phi_3} + 50\frac{h^2\omega_i}{\phi_3^2} + 96\frac{h^3\omega_i^2}{\phi_3^3}\right). \end{aligned}$$